

Killers on the loose

Across the developing world, seriously ill patients can't be sure whether they're purchasing life-saving medicines or worthless dummy pills. This scandal demands a stronger response from aid donors, governments and the drugs industry.

If you have ever been to southeast Asia, you'll probably have visited a street market laden with counterfeit goods. Maybe you even bought a knocked-off 'Rolex' watch or some pirated DVDs. It's all part of the Asian experience. Sure, it's illegal. But many visitors view the counterfeiters as Robin Hood-style outlaws, robbing rich Western companies to bolster the local cash economy.

But there is another side to Asia's culture of counterfeiting, which tourists don't usually get to see. The continent's pharmacies are awash with fake drugs that contain little or none of the labelled active ingredient. Produced mostly in China and India, these bogus products are also exported to Africa and the rest of the developing world. This trade is far from benign. It spawns chronic sickness, and leaves orphans and widows in its wake. It is run by ruthless individuals with scant regard for human life, and flourishes wherever they can induce corrupt officials to turn a blind eye.

Thankfully, awareness of the developing world's problems with fake pharmaceuticals is growing. Scientists and doctors who are trying to improve standards of healthcare in some of the world's poorest countries have helped to raise the alarm (see page 132). And in recent years, several programmes have been established to address the issue. The International Chamber of Commerce, for instance, launched a Counterfeit Pharmaceuticals Initiative in 2003 that acts as a clearing-house for information about fake drugs.

Multinational drug companies have also begun to get involved. Several have come together to form the Pharmaceutical Security Institute, which gathers information on counterfeiting so that prosecutions can be pursued. Such actions have been driven by a growing

Internet trade in counterfeit drugs that is hitting 'big pharma' where it hurts — by targeting their rich, developed-world consumers.

But the industry, wary of sharing commercially sensitive information and fearing that publicity could undermine consumer confidence in its products, is still playing its cards close to its chest. And it is less than fully engaged with the problem in the developing world, where the human consequences of drug counterfeiting are most severe. Organizations fighting the menace of fake drugs say they'd like to do more but don't have the resources. Governments and drug companies should put those resources in place, and exert greater pressure on countries that harbour counterfeiters to bring them to justice.

One official working for a Western agency was told to lay off the issue, after invoking the ire of senior Cambodian officials by drawing attention to the country's fake drug problem and the corrupt environment in which it flourishes. Given that aid projects can only take place with the blessing of host governments, his superiors were presumably worried about jeopardizing all their efforts.

Such attitudes are misguided and unforgiveable. Many of the people behind the trade in fake drugs are former narcotics smugglers who have switched from a business where policing is stringent and penalties severe. These vicious, organized criminals are literally getting away with murder — a situation that should not be tolerated.

There are no easy answers. In the countries where the problems with fake drugs are most severe, it can be hard to tell who is part of the solution and who is part of the problem. But this is no excuse for inaction. As the philosopher Edmund Burke observed: "All that is necessary for the triumph of evil is that good men do nothing." ■

Bad faith at Los Alamos

The breakdown of an old contract threatens to leave a great national laboratory gravely weakened.

For an outsider, it's hard to imagine that conditions at Los Alamos National Laboratory in New Mexico could get much worse. When two disks containing classified data went missing last summer, federal investigators descended on the lab, operations were suspended for months, and scientists were berated as "buttheads" by Peter Nanos, the lab's director. Now, it turns out, those disks never existed (see *Nature* 433, 447; 2005).

Incredibly, Los Alamos could slide still further downhill. Its long-standing contract with the University of California, which manages the lab for the US Department of Energy, will be opened up to competition. At stake for lab employees is the academic clout that comes with the university's reputation, as well as a sterling pension plan for which many at the ageing lab are already eligible. Unless these benefits are protected, insiders fear that scientists will flee in droves.

Since Los Alamos was founded in 1943, its relationship with the university has been good for scientists. They enjoy the same privileges as university employees, including a lower tuition rate to the university's campuses and membership of the retirement programme.

But since 1999, university officials have spent countless hours responding to scandals, so some wonder whether it is worth the

\$9-million fee that the university receives. And the energy department has been embarrassed again and again by a never-ending string of security incidents — many of which may have been inflated by political animus towards the laboratory and its managers.

Looking for others interested in running the lab isn't a bad idea, but the exercise is creating too many uncertainties for scientists. If the lab were to go to another company, the university's pension plan and other benefits could be lost for ever. Along with the director's penchant for name-calling, it is no wonder that many older scientists are reportedly taking early retirement.

The best hope for fixing the contract might be Samuel Bodman, the incoming energy secretary. Bodman's credentials as former professor of chemical engineering at the Massachusetts Institute of Technology played well with the lab's rank-and-file during a meeting in February. And his previous job as president of Fidelity Investments has won him praise from congressional overseers. If he chooses to, he may be able to rebuild the lab's broken contract with the university, restore trust between the wounded parties, and ensure that researchers' benefits are protected. If he doesn't, Los Alamos could become just another ghost town in the New Mexico desert. ■





Birthday bash
Accusations hit
academy's
celebrations
p126



Youth movement
Europe charts
fresh course for
training in research
p127



What a meth!
Home drug labs
prove to be toxic
time bombs
p129



Monkey puzzle
Researchers set to
figure out marmoset
genome sequence
p130

NASA's funding shortfall means journey's end for Voyager probes

Tony Reichhardt, Washington

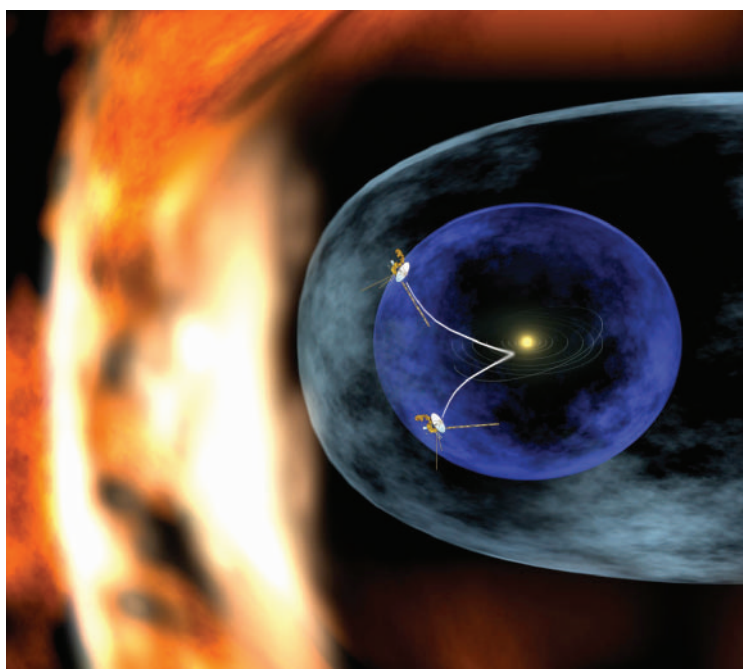
NASA has told scientists working on some of the agency's longest-running space missions — including the twin Voyagers now speeding towards the edge of the Solar System — that they may have to shut down operations in October to save money.

The decision — which NASA officials say is not yet final — has angered scientists, who call it penny-wise and pound-foolish, and say that it is being made without the usual formal review by the science community.

The agency's Earth-Sun System division originally planned to spend \$74 million next year to operate spacecraft that study the Sun's environment and to pay for data analysis. But the budget has since been cut to \$53 million, as NASA struggles to fund all of its planned projects. Faced with the shortfall, division officials last month informed the managers of seven missions that are past their prime lifetime — Voyager, Ulysses, Polar, Wind, Geotail, FAST (Fast Auroral SnapshoT) and TRACE (Transition Region and Coronal Explorer) — that there is no money to keep them operating after the current fiscal year ends in October.

Launched in 1977, Voyagers 1 and 2 are now more than 14 billion and 11 billion kilometres from Earth, respectively. Having visited all the outer planets except Pluto, they are on their final quest — to locate the shifting boundary between the Sun's domain and the realm where interstellar space begins. Ground antennas are in daily contact with the spacecraft, which are expected to last until about 2020 before giving out. The Voyagers cost NASA about \$4.2 million a year for operations and data analysis.

To afford new projects, NASA occasionally has to turn off spacecraft that are still



On the edge: the Voyager spacecraft are on the verge of interstellar space.

working but that have exceeded their life expectancy. Every two or three years, a 'senior review' of outside scientists ranks the science value of currently operating missions to help the agency plan which ones should be extended and which shut down. The last such review for Earth-Sun System missions was in August 2003, and the projects now facing termination ranked lowest in that exercise.

Out of time

But the panel never suggested that the low-ranking missions should be shut down this year. For example, Ulysses, launched in 1990 to explore the Sun's polar regions for the first time, was recommended to continue until 2008, and Wind until 2007. The Voyagers were recommended for funding until at least 2006, after which they would have to be reviewed again.

Stephen Keil of the National Solar Observatory, who chaired the 2003 review, says his panel assumed a budget of \$74 million for 2006, and although their rankings might have been the same given a lower budget, their

advice on how to allocate resources would probably have been different. Voyager project scientist Edward Stone of the California Institute of Technology says that instead of holding another review (the next one is scheduled for spring 2006), NASA "just worked up the old list until they matched the lower budget. But that really defeats the purpose of the senior review, which at the time was predicated on a certain level of funding."

Lennard Fisk, a University of Michigan space scientist who chairs the National Academy of Sciences Space Studies Board and is a former head of NASA space science, calls the threatened cuts "an extremely foolish thing to do". Voyager, he says, is entering one of the most

interesting scientific phases of its long life as its particle detectors approach the edge of the Solar System. "There is no replacement for Voyager," he adds — no other working spacecraft have travelled so far from Earth, and no new ones are being built to follow them. Similarly, says Fisk, it "doesn't make sense" to turn off Ulysses just as the Sun comes to the end of a 22-year magnetic cycle.

Scientists from the threatened missions are already lobbying their congressional delegations to try to get the money restored, and Fisk thinks NASA could easily solve the budget problem by shifting funds within its \$2-billion budget for the Earth-Sun System.

That may, in fact, happen. "There's been no final decision at NASA headquarters to terminate any of these missions, despite what budget figures may imply," says NASA spokeswoman Gretchen Cook-Anderson. But for now, project scientists say they have no choice but to take the threat seriously. Having been told by NASA that there is no money available after October, Stone says, "we are currently developing a plan for shutdown".

W. FEIMER/NASA

Senate resignations mar academy's birthday

Alison Abbott

The European Academy of Sciences and Arts in Salzburg, Austria, celebrated its 15th anniversary last weekend with considerable pomp and ceremony. But the party has been pooped by allegations of mismanagement.

The accusations centre on a lack of transparency about the body's finances and the alleged failure of its president, Salzburg heart surgeon Felix Unger, to properly abide by the academy's rules.

Unger denies the allegations, which were first reported in the German newspaper *Süddeutsche Zeitung* on 4 March. He says that they have been made up by people who have tried, but failed, to get money from the academy.

The academy has some 1,200 members, nearly half of them from Austria or Germany. It is divided into seven sections, ranging in scope from religion to science, and has several institutes under its umbrella. These include

two run by Unger himself — the European Heart Institute and the European Institute of Medicine — which share the same address.

The academy has the patronage of the city of Vienna and the Austrian government, and has a grant of €150,000 (US\$200,000) from the European Commission, primarily to help cover its administrative costs. Its website also features a long list of companies and organizations designated as partners, benefactors, sponsors and donors.

But in the weeks running up to the anniversary celebration, three members of the academy's 18-strong senate resigned, expressing many concerns including lack of clarity on how money flows into, and out of, the academy.

Justin Stagl, a sociologist from the University of Salzburg and former chairman of the academy's nominations committee, said in his 8 February resignation letter that he

had resigned because he was unable to persuade Unger to stop nominating new members without discussion with, or the consent of, the senate. He also noted that many ordinary members had resigned and had criticized a 'personality cult', 'clientelism' and 'empty boasting' at the academy.

Ernst Pöppel, a neuropsychologist at the University of Munich, was dean of the science section of the academy but says he resigned because Unger had become "a one-man show", changing the legal basis of the academy without discussion with the senate and not opening the accounts of the academy to scrutiny. "There is no transparency in the finances of the academy," he says. "For example, once a very large debt appeared, but was then wiped out, without anyone being able to find out why we had gone into debt and who had paid the debt off."

Bernd-Olaf Küppers, a philosopher from the University of Jena, who had been dean of the academy's humanities section, said that he too resigned on the grounds of its lack of financial transparency.

Unger says that the academy aspires to forge a European identity through interdisciplinary, transnational and bridge-building activities. He insists that the financial accounts have always been open to the senate. "It's all lies," he says.

Unger has his supporters in the senate. Eugen Biser, a theologian at the University of Munich and dean of the academy's world religion section, says that the accusations are born of "revenge and spite" and that the academy must allow the affair to blow over. Gilbert Fayl, foreign secretary at the academy, says the accusations are "purely personal". ■

Additional reporting by Tamara Grüner.



Cool location: Salzburg in Austria is home to the European Academy of Sciences and Arts.

Anthropologists walk tall after unearthing hominid

Rex Dalton, San Diego

The remains of a hominid that date back nearly 4 million years have been uncovered by a field team in Ethiopia.

The discovery, announced in Addis Ababa on 4 March, looks set to provide crucial new evidence on early man's ability to walk upright.

Palaeoanthropologist Yohannes Haile-Selassie and his colleagues found the bones near Mille Town, about 520 kilometres northeast of Addis Ababa. "It is too early to tell what species is represented," says Haile-Selassie, a curator at the Cleveland Museum of Natural History in Ohio.

A detailed description of the fossils will be submitted for publication once excavations and bone analysis are complete,

which may not be for another year.

By studying the bones and muscle connections, anthropologists can decipher the mechanics of how different hominid species came to walk as they evolved from primates.

In addition to bones of the legs, hip and spine, the team uncovered a complete scapula, or shoulder bone. It is extremely rare to secure a scapula, as the thin bone is typically damaged by the forces of time.

The shoulder bone should provide important new data on arm and body functions. The leg bones indicate that the individual hominid was taller than Lucy (*Australopithecus afarensis*), a 3.2-million-year-old fossil found in 1974 about 60 km south of the new site.

Examination of animal fossils found with the new hominid specimen suggests that it lived between 3.8 million and 4 million years ago.

The team's preliminary analysis indicates the hominid is a form of *Australopithecus*. It may be an *Australopithecus anamensis* — known only by a few bones and teeth found in Kenya, where that species lived about 4 million years ago.

The director of the Cleveland Museum, palaeoanthropologist Bruce Latimer, is a co-leader of the project, but he was not in the field when the fossils were found. Fossils on the surface were immediately collected by the team to prevent them from being damaged. Digging is expected to begin later this year as the team looks for other specimens. ■



Supporters of a new charter hope that it will give junior researchers badly needed leverage.

Career charter sets out rights of Europe's young scientists

Quirin Schiermeier, Munich

Ask a hundred young scientists in Europe for their views on their supervisors, funding agencies and university administrators, and you'll get a hundred different answers. But now a 'career charter' is to be published that, its backers say, will encourage higher standards for the treatment of junior researchers.

The European Commission, which intends to release the charter on 11 March, wants the 25 member states of the European Union (EU) to adopt it in the hope that this will help to introduce better and more consistent scientific training across the continent. But some researchers are already expressing doubt that it will make much difference.

The charter declares that all EU researchers, including PhD students, should be recognized as professionals and treated accordingly. It defines the rights and duties of both young researchers and their supervisors. And it sets out minimum standards for training, employment conditions and social-security coverage for fixed-term researchers working in universities, research institutes and industrial laboratories.

There is considerable demand for such standards. Many young researchers, especially those in eastern and southern Europe, are working under ill-defined terms of employment, without proper contracts, and are effectively at the mercy of their supervisors.

Young scientists' organizations, such as Eurodoc or the Marie Curie Fellowship Organization, are swamped with letters from researchers who have fallen out with their supervisors, or who feel discriminated against on sexual, racial and other grounds.

In January, for example, an Irish biochemistry postdoc working in Austria told Eurodoc that her supervisor had fired her after she had told him that she was pregnant.

Will the charter help her? Sceptics point out that the commission has no legal means to force EU members, or individual universities, to comply with it. But it will invite member states to report annually on their progress in adopting the charter. And supporters say that its very existence will give young researchers some useful leverage.

Christine Heller del Riego, coordinator of the working group on career development for Euroscience, a grass-roots research organization, sees the charter as a badly needed start. Del Riego, who is an electrical engineer at the University Pontificia Comillas in Madrid, is currently drafting a career development strategy for young researchers at her university. "If you can refer to an EU recommendation, then you have a leverage that might really get things moving," she says.

Spain's prime minister, José Luis Rodríguez Zapatero, promised last year to improve the statutory rights of young researchers, for example. But he has since ignored petitions from the FJI, the Spanish young researchers' federation, to go ahead with the promised changes. "Now the EU charter can be used as an additional pressure," del Riego says.

Others say the charter doesn't go far enough. Martin Grabert, head of the Brussels-based EU liaison office of German research organizations, says he is disappointed by what he terms its "timid, non-committal character", adding: "This is a typical EU approach, a mere minimal consensus with a good chance of falling flat." ■

Gene-therapy trials to restart following cancer risk review

Erika Check, Washington

Gene-therapy trials for children suffering from severe combined immunodeficiency disease (SCID) are set to resume in the United States, despite another cancer case in a French trial of the technique.

On 4 March, an advisory meeting in Rockville, Maryland, heard evidence that the cancer risk exposed by the French trial may be specific to one form of the disease, known as X-linked SCID.

The advisory panel recommended that a gene-therapy trial for another variant of the disease, called ADA-SCID, should be allowed to proceed.

The French trial has so far successfully treated six children with X-linked SCID, but three other patients have since developed cancer, and one of these has died. The third of the cancer cases was reported in January, and prompted the US Food and Drug Administration to suspend three US gene-therapy trials on SCID (see *Nature* 433, 561; 2005).

The leader of the French trial, Alain Fischer of the Necker Hospital for Sick Children in Paris, has already said that he will not continue his trial until he has made changes to the methods he uses to deliver the therapeutic gene to children with X-SCID. But the US panel said that US trials on X-SCID could proceed on patients who have failed to respond to other treatments or are likely to do so.

And it said that, so far, the risk of cancer seems unique to X-SCID, and may not apply to ADA-SCID. At the meeting, oncologist Utpal Dave of the National Cancer Institute described unpublished research suggesting that some of the risk of cancer in the X-SCID trials was related to that form of the disease, and might not appear in other variants of SCID.

Dave's data linked the gene that is missing in children with X-SCID to cancer in mice. He theorized that the method used in the French trial to replace this gene could be increasing the children's risk of cancer. But he found no data to suggest a similar risk for ADA-SCID.

"Although not conclusive, these are the first data that suggest the gene may play a role in these cancers," says Daniel Salomon, a molecular biologist at the Scripps Research Institute in La Jolla, California.

There is one planned US trial in ADA-SCID, led by Donald Kohn of the Childrens Hospital Los Angeles. "I'm glad they agreed that ADA clearly has a safer record to date," Kohn says. ■

Climatologists seek clear view of Asia's smog

David Cyranoski and Ichiko Fuyuno, Tokyo

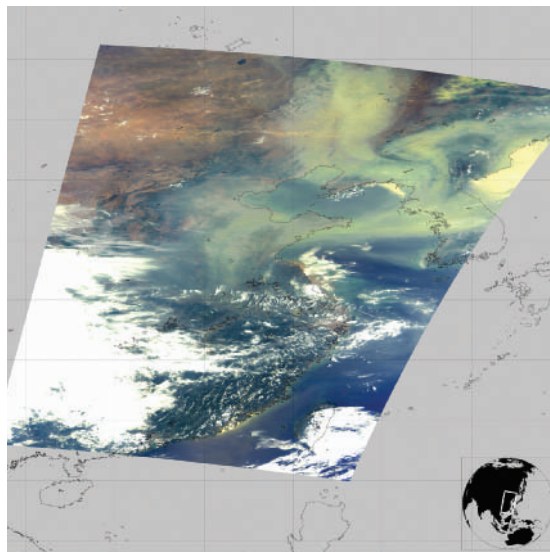
An international team gathered in South Korea this week to kick off a research project on Asia's persistent brown clouds of smog. The effort is the first long-term study into the effects of the haze on local and global climate.

The brown clouds frequently envelop vast portions of Asia. They wax and wane with the seasons but can be kilometres thick, causing respiratory disease and interrupting climate cycles.

Researchers on Project Atmospheric Brown Clouds, run by the United Nations Environment Programme (UNEP), are hoping that their study of the smog will help to spur preventative measures against the pollution. This could be especially important in the face of China's accelerating industrial boom.

The clouds consist of ash, acidic chemicals and carbon, which come from the burning of fossil fuels and wildfires. One of the most intensely polluted and most studied brown clouds, which hangs over the Indian Ocean, has recently been found to consist largely of the smoke from dried manure used for cooking (C. Venkataraman *et al. Science* 307, 1454–1456; 2005).

Previous research has shown that the haze over the Indian Ocean blocks up to 15% of the Sun's radiation (S. K. Satheesh and V. Ramanathan *Nature* 405, 60–63; 2000). Such clouds are thought to cause the ground to cool and the atmosphere to warm, which can affect monsoons and other rainfall patterns. Particulate matter can also alter rain formation. But little is known about the details of how the clouds change over time, or how this alters the climate, says Veerabhadran



Foggy notion: Asia's hazy smog comprises fumes from forest fires (yellow) and pollution from populated areas (white).

Ramanathan, an atmospheric scientist at the Scripps Institution of Oceanography in La Jolla, California, who is involved with the UNEP project. "We need to watch the whole seasonal cycle for several years," he says. "There have been many theories about how this affects global climate. We have to observe it."

The UNEP project aims to follow the Asian clouds for at least five years; funding of US\$10 million has been secured for this period from the US National Oceanic and Atmospheric Administration and the National Science Foundation. The programme has two main observatories: one in the Maldives and one on the Korean island of Jeju. Each is stocked with more than 30 instruments that can measure incoming sunlight, temperature and the chemical

composition of the clouds. Tens of smaller labs throughout the region will also contribute data.

The scientists involved say they are particularly concerned that the industrial boom and desertification currently going on in China are throwing ever more smog ingredients into the atmosphere. "East Asia might have more surprises than India — it is of a greater concern," says Ramanathan.

But the attempt to study China's impact on the clouds could become politically sensitive. Japanese scientists estimate that some 30% of the soot falling on their country comes from China, whereas Chinese researchers place the estimate as low as 5%. "There is no consensus," says Yuanhang Zhang, an atmospheric chemist at Peking University who is representing China in the UNEP project.

Shiro Hatakeyama, a specialist in atmospheric chemistry at the National Institute for Environmental Studies in Tsukuba, north of Tokyo, says that China has not provided sufficient information about its emissions to feed into models that predict the effects of the smog. "So far, China's efforts are not satisfactory," says Hatakeyama.

Zhang admits that there is a lack of data for most pollutants other than sulphur dioxide in China, making it hard to assess their contribution to the haze. Hatakeyama hopes that the project will put pressure on China to collect and provide more information.

That may already be happening. Zhang says that there are plans to set up about ten observatories in the country to contribute to the programme — a meeting is scheduled for April to work out the details. ■

Biosafety law brings stem-cell research to Brazil

Laura Nelson

Brazil has passed a landmark biosafety law that will legalize human embryonic stem-cell research, and may open the way for the cultivation of genetically modified crops.

The move ends a period of uncertainty and frustration for Brazilian cell biologists by allowing them to use stem cells from frozen embryos left over from *in vitro* fertilization. It also establishes a clear process for the approval of genetically modified crops.

Although the law was vigorously opposed by church groups in Brazil, it was passed by the lower house of Congress on 2 March and is expected to be signed into law by President Luiz Inácio Lula da Silva in the next few weeks.

Brazilian stem-cell researchers welcomed the law — even though it forbids the cloning of stem cells for therapeutic use. "We are so happy," says Lygia Pereira, a cell biologist at the University of São Paulo.

A year ago, a preliminary version of the law passed by the lower house in Congress blocked the use of embryonic stem cells for research. But scientists and patient groups fought back to win the support of the Senate last October, and then that of the lower house this month.

"It is a huge step forward from the first draft that prohibited any kind of research with human embryos," says Pereira.

Previous biosafety regulations hadn't covered the use of stem cells, focusing

instead on transgenic crops, banning their planting and commercialization. But genetically modified seed is being smuggled into Brazil, and the new law may allow for the crops' official approval.

Under the law, a body called the National Technical Commission on Biosafety (CTNBio) will advise on whether a genetically modified crop can be planted and commercialized. A council of government ministers will then have the final say.

Leila Oda, president of the National Biosafety Association, says that the law will facilitate research in agricultural biotechnology in Brazil. "Anyone wanting to do research simply applies to CTNBio," she says. ■



It's a wrap: a Tennessee cop wrestles with the remains of a home lab for making methamphetamine.

Police urge speedy action to clean up home drug factories

Emma Marris, Washington

Abuse of methamphetamine is creeping across the United States from the west coast, producing a welter of toxic home labs, a congressional committee heard last week.

The drug, a cheap, potent and highly addictive stimulant, can be 'cooked' on a hotplate from easily acquired ingredients such as cold remedies and fertilizers. And the mess left behind can affect the health of police who raid the premises, children who live there, and even subsequent tenants.

"During the cook, it reaches levels that can be extremely toxic," says John Martyny, a toxicologist at the National Jewish Medical and Research Center in Denver, Colorado, who testified at the hearing.

After complaints from police in his district east of Nashville, where 'meth' use has boomed in a big way, congressman Bart Gordon (Democrat, Tennessee) put together a bill to fund research into the drug's health risks and to devise standards for decontamination. The bill's bipartisan supporters expect it to move speedily into law.

If enacted, the bill will charge the Environmental Protection Agency (EPA) and the National Institute of Standards and Technology with researching the health risks of meth labs and setting standards for decontamination, as well as funding the development of better field-test kits.

Meth users who make their own drug do so wherever they can: at home, in motels, even in cars. Various recipes exist — most are available on the Internet — but even the simplest are dangerous and toxic. Typically, ephedrine is separated from a decongestant using solvents, then added to chemicals such

as ammonia, hydrochloric acid, phosphorus or iodine. The resulting methamphetamine is a fine particulate that gets everywhere. And many of the chemicals involved are absorbed by porous surfaces such as carpets, curtains, wallpaper and clothes.

Sheriff David Andrews of Putnam County, Tennessee, says he hates entering meth labs. "Every time I go around one, I get a headache, and I have a detective who breaks out in red blotches," he says. Setting standards would not only help law enforcement, Andrews points out, but also mortgage lenders who, without a decontamination standard, must take a financial loss when they repossess a house that has been used as a meth lab.

Julie Mazzuca is president of the Meth Lab Cleanup Company based in Post Falls, Idaho, where business is booming. She and her crew charge a few thousand dollars to don respirators and scrub away traces of methamphetamine from meth labs. She has seen the health effects at first hand. "You see respiratory problems, throat irritation, skin irritation and nausea," she says.

Meth sites, Mazzuca says, are unbelievably filthy. "On one of my first clean-ups I removed seven tonnes of residue from a mobile home," she says.

Although anecdotal evidence shows that meth labs are unhealthy, there is little research on the subject. Some US states have standards for detectable levels of methamphetamine, but they are based on the detection technology rather than knowledge of the health risks. "There's a lot of seat-of-the-pants stuff going on with these clean-ups," says Charles Salocks, a toxicologist with the California EPA. ■

Science agencies get fresh paymasters in Republican revamp

Geoff Brumfiel, Washington

NASA and the National Science Foundation (NSF) will answer to a new set of masters in Congress, following a wide-ranging shake-up of the committees that set their budgets.

In changes pushed through by the Republican leadership in both the Senate and the House of Representatives, the appropriations subcommittees that dealt with the two science agencies — as well as a range of other programmes, from housing to veterans' affairs — have been abolished.

NASA and the NSF will now be overseen by new subcommittees, where they will compete for funds with the Department of Commerce, the criminal-justice system and the state department.

It isn't yet clear what the change will mean for the agencies' budgets. But it poses a challenge to agency chiefs, who now have to explain what they do to a new set of people in Congress, beginning at hearings this week. "Change is good," says Arden Bement, director of the NSF, adding caustically that "sometimes, no change would be even better".

The House and Senate subcommittees are now chaired by, respectively, Frank Wolf (Republican, Virginia) and Richard Shelby (Republican, Alabama). Shelby crossed swords with scientific leaders in 1999 when he championed a bill to open up scientists' data for scrutiny by industry groups (see *Nature* 397, 459; 1999).

But many of the new committees' members have had scant dealings with science agencies in the past. "We have a whole new group of people who need to be educated about the importance of the NSF and NASA," says Tobin Smith of the congressional affairs office at the Washington-based Association of American Universities.

Each year, the subcommittees take the president's budget request and convert it into the final budgets received by agencies and government departments. The process is heavily dependent on the personalities involved, and the NSF, in particular, has benefited in recent years from allies on the old subcommittees who sometimes gave it more money than the White House requested.

Both Shelby and Barbara Mikulski (Democrat, Maryland), the top Democrat on the Senate committee, have major NASA facilities in their states, which should ensure strong support for the space agency, some lobbyists note. ■

German prize winner misses ceremony after misconduct allegation

Munich One of the ten winners of Germany's most prestigious scientific prize was absent from the prize-giving ceremony last week in Berlin. The DFG, the country's main funding agency, withdrew its invitation to Stefanie Dimmeler, a molecular cardiologist at the University of Frankfurt, after anonymous accusations of scientific misconduct.

Dimmeler, who studies atherosclerosis, was selected in December for the DFG's €1.5-million (US\$2-million) Leibniz Prize. The agency subsequently received letters pointing to identical figures in at least three of her publications. Dimmeler admitted a "mistake" in a paper published in *Nature Medicine* in 2003, but says she informed the journal, which published a correction in September 2004.

The DFG says it will decide whether to award the prize money after an investigation by the University of Frankfurt. The DFG's ombudsman for scientific misconduct, Hans-Heinrich Trute, says Dimmeler has been open about errors in her publications, but the allegations must be fully investigated.

Mine hosts detector to put neutrinos to the flavour test

Washington Scientists at Fermilab in Chicago have begun an experiment that they hope will help them to understand the properties of one of the most mysterious subatomic particles, the neutrino.

Neutrinos are generated in some nuclear reactions, rarely interact with any kind of matter, and come in three different types, or 'flavours'. In 1998, researchers unexpectedly found evidence that neutrinos from the Sun may spontaneously switch from one flavour to another — a fact not predicted by the standard model of particle physics.

To study this flavour switching, Fermilab physicists have constructed a \$170-million neutrino beam and detector. The project, known as MINOS, will hurl trillions of neutrinos each year through 735 km of solid rock to a detector in the Soudan iron mine in Minnesota. Just a few thousand neutrinos will be detected, but scientists hope this will be enough to catch flavour switching in action. They believe the study may ultimately allow them to revise the standard model.

Parliament plans move to make 'European MIT'

Strasbourg A cross-party group of 130 European politicians has proposed turning the parliament building in Strasbourg into a European Institute of Technology (EIT).

Marmoset heads queue for genome sequencing

Washington Twelve more organisms have been selected for genome sequencing by the US National Human Genome Research Institute (NHGRI) in Bethesda, Maryland — all chosen for the insights they can provide into human genetics. The marmoset (*Callithrix jacchus*, pictured), is the first New World monkey to get the treatment. The list also includes the skate (*Raja erinacea*), which sits at a pivotal node of the evolutionary tree, and the insect *Rhodnius prolixus*, which carries the parasite that causes Chagas' disease.

The other species are the sea slug *Aplysia californica*, the pea aphid *Acyrtosiphon pisum*, the insects *Nasonia vitripennis*, *N. giraulti* and *N. longicornis*, the soil amoeba *Acanthamoeba castellanii*, and the fungi *Schizosaccharomyces octosporus*, *S. japonicus* and *Batrachochytrium dendrobatidis*.

Don't expect the results overnight, warns a



spokesman for the NHGRI, which does a large chunk of the world's public sequencing. "It takes a while, once an organism is approved, to find the right strain or the appropriate sample, and then to make libraries and do some preliminary testing." The NHGRI adds species to its list two to three times a year.

The idea of creating the EIT, modelled on the Massachusetts Institute of Technology, was put forward last month by European Commission president José Manuel Barroso as part of his strategy to make Europe more competitive.

A parliamentary group known as the Campaign for Parliament Reform leapt on the idea. The group campaigns for an end to



Strasbourg: a chance to swap politics for science.

the notorious 'travelling circus' that occurs when the entire European Parliament moves from Brussels to Strasbourg for a few days every month to hold some sessions.

Abandoning the Strasbourg building would save €200 million (US\$260 million) every year in travel costs, the group says, but is politically sensitive. It hopes Strasbourg might accept the EIT as adequate compensation for losing the parliament.

India builds up research base with two institutes

New Delhi The Indian government is to create two new research institutes and establish an autonomous funding agency under the aegis of the prime minister, Manmohan Singh.

The chairman of the prime minister's science advisory council, C. N. R. Rao, says he has been pushing the scheme for more than a decade. "I am happy that it has finally

come through," he told *Nature*. The agency, the National Science and Engineering Research Foundation, is to be modelled on the US National Science Foundation and will have an initial annual budget of Rs10 billion (US\$230 million). The two new institutes — in Pune and Kolkata — will each cost Rs5 billion and be modelled on the Indian Institute of Science in Bangalore.

In another effort to build up basic science, research funding in some of India's seven institutes of technology will be increased.

Environment agency gets scientist as head

Washington For the first time in its 35-year history, the US Environmental Protection Agency (EPA) is to be run by a scientist.

Stephen Johnson, a pathologist by training, was last week nominated by President George Bush as the successor to Mike Leavitt, who left the EPA for the Department of Health and Human Services in January.

Johnson, a career employee at the EPA, had been acting head of the agency since Leavitt's departure. He ran the agency's pesticide programme before becoming deputy administrator in 2003.

Bush has come under fire for opposing environmental regulation, and his first EPA administrator, Christine Todd Whitman, resigned in 2003. She has since criticized the Bush administration for being too ideological.

Even groups usually critical of the Bush administration cautiously welcomed the choice. "We're hopeful that Johnson's nomination means that the administration will make decisions based on science and protective of public health and the environment," said the US Public Interest Research Group in a statement.

Black market:
illicit medical
supplies for sale
in Phnom Penh.



Murder by medicine

Across the developing world, people are dying after being peddled fake pharmaceuticals. Peter Aldhous reports from southeast Asia, where scientists, doctors and regulators battle against organized crime.

Sam Veasna's loss is still keenly felt in Cambodia, five years after his untimely death. In a desperately poor country where most people are preoccupied with the daily struggle for survival, he was a tenacious field biologist and passionate advocate for conservation whose work made a real difference — until he fell prey to malaria.

In November 1999, Veasna ventured to the remote north of the country in search of the elusive kouprey — a wild cow that is Cambodia's national animal, but which may already be extinct. True to form, the kouprey evaded Veasna. But the mosquitoes didn't. Back home in Siem Reap, where he ran the provincial government's wildlife office, Veasna fell ill. He was treated with tablets labelled as mefloquine, an antimalarial drug, which should have restored him to health. But his fever continued to rage.

"By the time we realized what was going on, he was slipping into a coma," recalls Colin Poole, who heads the Asia programme of the

Wildlife Conservation Society, based in New York. "We organized a medical evacuation to Bangkok." But it was too late. Veasna was pronounced dead on 3 December, shortly before the plane that was to whisk him to Thailand touched down in Siem Reap. He was just 33, and left behind a wife and three children.

Veasna was almost certainly a victim of the trade in counterfeit drugs — bogus products that often contain no active ingredients, or the wrong ones. Cambodia can ill afford to see its brightest lights snuffed out in this way, having already lost a generation of intellectuals in the 1970s to the genocidal Khmer Rouge regime. Yet a criminal investigation into Veasna's death soon collapsed through lack of evidence — the offending tablets had been thrown away. If he hadn't been internationally renowned, things would probably never even have got that far. Many ordinary Cambodians are thought to suffer a similar fate each year, their deaths gaining scant media attention and prompting no official response.

Hologram wars

In his office at the Hospital for Tropical Diseases in Ho Chi Minh City, Vietnam, Jeremy Farrar tosses me two blister packs of tablets on which the name 'Guilin Pharma' is printed in blue. The company's factory in Guangxi province, southern China, is the leading source of artesunate, a highly effective treatment against malaria.

One of the packs doesn't look quite right. The printing seems slightly washed out. And, tellingly, its security sticker is revealed on close scrutiny not to be a hologram, but a piece of shiny foil with a printed pattern. After examining the two, I'd be fairly confident of spotting the fake if I was offered it over a pharmacy counter.

Unfortunately, the fake artesunate circulating throughout southeast Asia today is much harder to detect. Paul Newton, who is studying the problem of counterfeit drugs in the region from

T. PAGE/CORBIS

No one knows how many deaths in the developing world can be blamed on fake pharmaceuticals. But thanks to the diligence of researchers who are monitoring the effectiveness of malaria treatments, we now do know that the trade in counterfeit antimalarial drugs is rife across southeast Asia. Scientists are battling the menace of fake drugs by raising awareness among doctors and their patients, and running lab tests to screen for counterfeits. But there's only so much they can do. Next week in Paris, at the 2nd Global Forum on Pharmaceutical Anticounterfeiting, experts will try to plot the way forward. It's already clear that this will require more rigorous efforts to crack down on the ruthless criminals who profit from the trade in bogus drugs.

Innocent victims

The alarm about fake antimalarials began to sound just months before Veasna's death. In 1999, Jan Rozendaal, who then ran the European Union's Cambodia Malaria Control Project, began to hear reports of patients who failed to respond to treatment with artesunate. This drug, derived from a Chinese herb, is highly effective against forms of the malaria parasite that resist earlier generations of antimalarials — particularly when used in combination with mefloquine. So if the parasite was evolving resistance to artesunate, it would have been a major blow.

But the problem wasn't drug resistance. Those who failed to respond to artesunate had bought suspiciously cheap supplies from small commercial pharmacies. So Rozendaal collected his own samples from similar outlets, and sent them for analysis to the Thai Ministry of Public Health¹. The drugs were sold in blister packs bearing the name of Guilin Pharma, a company in China's Guangxi province. But about half the packs

contained dummy pills with no active ingredient. Close inspection revealed subtle differences in the packaging that distinguished them from the genuine Guilin product. It was a similar story for bottles of mefloquine tablets, purportedly manufactured in Australia. In this case, the fakes had carved out a dominant market share.

Disturbed by his findings, Rozendaal contacted Nick White, a leading malaria researcher at the UK Wellcome Trust's unit at Mahidol University in Bangkok. The trust has units and collaborators across southeast Asia, so White was able to extend Rozendaal's investigation to cover the entire region. Over a year, ending in August 2000, volunteers purchased artesunate in Cambodia, Laos, Myanmar, Thailand and Vietnam. Of a total of 104 samples bought more or less at random, 38% were found to be fakes containing no active ingredient².

The researchers hoped their survey would raise awareness of the problem, and they subsequently called for international action to clamp down on the counterfeiters³. But a follow-up study, which concluded in February 2003, revealed that the situation had actually worsened. This time, 53% of 188 artesunate samples were useless fakes⁴. Worse, the counterfeiters had upped their game, improving their mimicry of the Guilin packaging to the point that the fake packs were virtually indistinguishable (see 'Hologram wars', below).

Back in 2000, many pharmacists probably knew they were selling knock-offs, says Arjen Dondorp, a clinician at the Wellcome Trust's Bangkok unit, who coordinated



Cambodia's health agencies are running awareness campaigns on fake drugs.

the second survey. "They'd say: 'This is a little bit less effective, but it's cheaper.'" In Cambodia, where the average annual income is about US\$300 and a course of artesunate sells for about US\$1.50, that was an especially compelling sales pitch. Today, Dondorp suspects that some pharmacists are being duped as well, and the price difference between genuine and fake artesunate has all but disap-

peared. "It's clearly organized crime, and highly sophisticated," says Dondorp. Sophisticated, maybe, but callous in the extreme. "You can think of it as attempted murder," he adds.

Streetlife

In December, I visit the Cambodian capital, Phnom Penh, to get a feel for the environment in which this trade flourishes. There are many reminders that this is one of Asia's poorest countries: at one junction, people are picking through piles of garbage for anything that might be reused. I wander the streets, searching out the city's small pharmacies. Most consist of a single open-fronted room, separated from the street by a simple counter. Packages of drugs line the walls inside. At each, I turn up with a scrap of paper on which I've written "artesunate or mefloquine" or "tetracycline". The last of these drugs is an antibiotic deployed against a variety of bacterial infections; it is also used, together with quinine, to treat malaria.

For someone used to the controlled



Copycats: (from left) the hologram on a genuine pack of Guilin Pharma drugs, a silver foil copy, and a convincing fake.

his base at the Mahosot Hospital in Vientiane, Laos, has documented seven distinct types of fake 'Guilin' artesunate. The counterfeiters soon began to incorporate security stickers bearing a real hologram, albeit a crude example. And the holographic stickers on the latest fakes, if anything, look more professional than the genuine version. Only an expert could reliably tell the two apart.

Yet it doesn't have to be that way, according to David Pizzanelli, sales and marketing manager

with Light Impressions, a holography firm in Leatherhead, near London. "The main problem is that the genuine hologram is terrible," he says. Depicting the limestone hills that surround the city of Guilin, plus a few Chinese characters, it has little three-dimensional relief. A passable imitation could be knocked up by more than 200 workshops worldwide, says Pizzanelli. And at least 15 millimetres across, the sticker is way too small. Surveys for credit-card companies have shown that security holograms need to be at least

twice this size if people are not to be fooled by crude imitations.

The answer, argues Pizzanelli, is to incorporate a more complex, repeating hologram into the foil of the blister pack itself. There are fewer than a dozen companies worldwide who could supply such a security feature, making life extremely difficult for the counterfeiters. But this could require an investment running to millions of dollars. And without clear evidence of how much revenue is lost as a result of counterfeiting, it's hard for cautious drug-company executives to justify spending such sums.

Lucy Jiang, who is responsible for international affairs at Guilin, says the company is talking to the World Health Organization and other experts about ways to protect its product from counterfeiting. But it has no plans to introduce a much more sophisticated hologram. "The cost is so high that we cannot bear it," says Jiang. **P.A.**

In the line of fire

Dora Akunyili has spent the past four years facing down corruption and tackling Nigeria's rampant problems with fake drugs. This crusade has been phenomenally successful, but has placed Akunyili's life in danger. Peter Aldhous caught up with her on a recent trip to the United States.

Q&A

How did you come to your job of regulating Nigeria's drug supply?

I got sick, and was given £12,000 (US\$23,000) by my employers for surgery in London. When I got there, it was diagnosed as irritable bowel syndrome, and I didn't need the surgery. So I returned the money. People said: "What type of woman is this? This is rare." Later, Nigeria's President Olusegun Obasanjo heard about this, and asked me about NAFDAC, the National Agency for Food and Drug Administration and Control. I said: "There is no system that cannot be corrected, if there is someone who is determined." After that, he gave me the job of director-general.

Did you have any personal experience with counterfeit drugs?

In Nigeria, there is hardly any family that does not have a history of somebody dying of fake drugs. My youngest sister died of diabetes in 1988. I'm a pharmacologist. I know it was fake insulin.

How bad were things when you took on your post?

I did not know the enormity of the problem until I came on board in April 2001. There was a total failure of the regulatory process. Fake and substandard drugs flooded Nigeria. Multinational drug companies divested and left. Our local drug manufacturers were closing up shop. How could they compete with somebody who is compressing chalk to make tablets, or adding olive oil into capsules and calling it multivitamins? People who had money went abroad for treatment. But the poor people were at the mercy of these criminals.

What did you do to tackle these problems?

We started with public enlightenment. Every two weeks, we publish the differences

between fake and genuine drugs in Nigerian newspapers. We also use the radio.

Most of the fake drugs in Nigeria come from India and China. In these countries, we established independent analysts to test drugs and give certification papers. Then we beefed up surveillance at all ports of entry into Nigeria. Now, if an aircraft arrives and we find fake drugs, it is grounded.

We have also devised systematic surveillance of markets and shops. We closed Aba market, in the southeast, for six months in 2003 for selling large quantities of unregistered, fake and substandard drugs. It's a serious thing to close a market. They apologized on radio and TV. Now they are fighting among themselves to keep it clean — if anybody makes NAFDAC close this market again, that person would be killed.

How many lives have been saved through your work?

Many deaths are not reported in Nigeria, and our people, culturally, are not into autopsy. But it must be millions. In 2004, we found that the incidence of unregistered drugs had fallen by 80% over two years.

How have the criminals responded?

They fought back! They put fetish objects in my offices. They put blood on feathers. I guess they were doing that to intimidate me. Then they started phone calls. They called my husband and warned: "Your family can easily be wiped out." My son, who was in high school in Benin, was clever enough to tell the students that I was not his mother. But that really pained me.

In August 2002, they went to our lab in Lagos and totally vandalized everything. I don't know what they used — maybe iron bars. The policeman on duty, they cut him to

"One of the drug barons arranged to assassinate me"



Before being brought in to head the National Agency for Food and Drug Administration and Control in April 2001, Dora Akunyili was a pharmacologist at the University of Nigeria Nsukka. In the 1990s, she worked for the Petroleum Special Trust Fund, which distributed government oil revenues for infrastructure projects.

pieces with machetes. In March 2004, they set our headquarters ablaze and, a few days later, they burnt down our labs in Kaduna.

One of the big barons arranged to assassinate me in my hometown in December 2003, as we were driving through. At first, I thought the sound was some Christmas fireworks, until our back windscreen shattered. A bullet went through my scarf, grazed my scalp, and went out through the front windscreen. I was crying, and the policeman in front of my car was on a walkie-talkie, telling my driver to keep going. A bus driver who was trying to block the shooters had his bus riddled with bullets. The man died on the spot.

Are you tempted to resign?

If I leave, it means that these criminals have triumphed. This, coupled with the trust that the president has in me, makes me feel that I have to struggle to the end. I'll be five years in the job in April 2006. By the grace of God, if I'm still alive, I will not go for renewal.

Will someone continue your work?

I really pray, and I'm also training some people. But this is a political position. My candidate may or may not be accepted. ■

Peter Aldhous is *Nature's* chief news & features editor.



Mean streets: the impoverished population of Phnom Penh makes an easy target for criminals selling fake pharmaceuticals.

environment of a typical Western pharmacy, this shopping trip is an eye-opening experience. Over two days, I visit a total of 27 outlets. Only once am I asked about the condition the drugs are supposed to treat. "It's for a friend with a fever," I say, before enquiring about the dose. The pharmacist says that my friend should take three tablets of tetracycline each day. But for how long? She scrutinizes the Khmer language information sheet supplied with the antibiotic, which doesn't seem to help. "As long as you like," she eventually suggests — vague advice that could place patients at risk, and promote the spread of drug resistance.

Faking it

As I pocket the packages, I contemplate the odds of them being genuine. Two official surveys conducted by the Cambodian Ministry of Health, with the backing of the World Health Organization, have shed some light on this question. National drug-testing labs in Phnom Penh and Bangkok looked at 230 samples of 24 pharmaceuticals purchased on the Cambodian market in 2000, including antibiotics and painkillers. They found about 3.5% of them contained less than 60% of the labelled quantity of active ingredient⁵. When the survey was repeated for a wider range of drugs purchased in 2003, 11% of the samples fell into this category⁶. Both surveys also turned up a small number of products that contained the wrong ingredient.

Some of these problems may represent shoddy manufacturing, rather than conscious

criminality. But ominously, the team that ran the second survey was unable to find any official record of the existence of five of the companies whose drugs they found on sale in Cambodia. 'Brainy Pharmaceutical', for instance, is supposedly based in Thailand, but is not registered with the appropriate authority in Bangkok. Its tested products included fake antibiotics with no active ingredients.

Chroeng Sokhan, vice-director of the Cambodian health ministry's Department of Drugs and Food, is trying to prevent bogus manufacturers from operating quite so brazenly in his country. His department, based in a run-down building in a shabby side street of Phnom Penh, has held workshops to educate retailers, and distributed posters warning of the dangers posed by counterfeit drugs — including some with photographs of known fakes. "Those products are harder to find on the market now," says Sokhan. In recent weeks, the anti-counterfeiting campaign has extended to broadcasting adverts on state-run television.

These activities have relied

This 'minilab' allows users to check whether drugs are genuine.



heavily on foreign aid, as do most efforts to combat drug counterfeiting in Cambodia. The US Pharmacopeia, a non-governmental body that develops standards to ensure drug quality, is working in the country with funding from the US Agency for International Development. Focusing on malaria drugs, it has already turned up samples of Brainy-branded quinine that — like the bogus company's antibiotics — don't contain any of the correct active ingredient.

Part of the strategy has been to involve provincial officials in testing drug samples, so that they take ownership of the problem. "They were very surprised when they found fake drugs in their provinces," says Lon Chantap, a researcher at the National Center for Parasitology, Entomology and Malaria Control in Phnom Penh, who works on the project.

Testing the quality of drugs is usually the preserve of large, well-funded national laboratories. But the US Pharmacopeia's project is also using a 'minilab'

developed by the German Pharma Health Fund, a charity financed by that country's drug industry.

This portable kit allows local technicians to analyse pharmaceutical ingredients using simplified versions of

techniques such as chromatography⁷. At the US Centers for Disease Control and Prevention in Atlanta, Georgia, analytical chemist Michael Green is similarly working on cheap drug-testing methods, including a red dye that turns yellow in the presence of artesunate⁸.

But in Cambodia, problems thrown up by grinding poverty run deeper than the ability to afford good drug-quality tests. Qualified staff are few in number and poorly paid. Sokhan can call upon just 13 national inspectors to police the country's entire drug supply. As the sun sets, he takes me for a drive through the streets that surround Phnom Penh's Olympic Market, which are packed with small pharmacies and wholesalers that sell drugs on to retailers in the provinces. Many, says Sokhan, are unlicensed outlets, and their continued existence is obviously a major source of frustration for him. "Unbelievable," Sokhan mutters, as we turn back towards the main road.

The next morning I continue my shopping trip. My final haul consists of five blister packs of Guilin-branded artesunate, some loose capsules of mefloquine, and five samples of tetracycline. In addition, I have three packets of a product called Malarine, which includes both artesunate and mefloquine pills and is distributed by Population Services International. This nonprofit body in Washington DC tackles problems with health and family planning in the developing world by placing affordable products into the commercial sector. At least these should be OK, I muse, before sending the samples to Green for testing.

Clean sweep

All of my drugs turn out to contain roughly the right quantities of active ingredients, and so seem to be genuine. Given that some of the pharmacies I visited carried Sokhan's posters, perhaps the campaign is having some effect. But Sokhan and other experts warn against reading too much into these results. As a wealthy Westerner, my sources suggest, I am more likely to be sold genuine drugs than a typical Cambodian customer. Perhaps if Veasna had been one of the many tourists who flock to Siem Reap to visit the ancient temples of Angkor, he would still be alive today.

For anyone familiar with Western practices of drug regulation, the solutions to Cambodia's problems with fake drugs seem clear: get police and customs officials to crack down on the distribution chain, allow only licensed pharmacists to continue operating, and close



Life or death? A selection of drugs (below) bought from street pharmacies in Phnom Penh.



any pharmacy caught selling counterfeit products. But Sokhan says that his department doesn't have the power to enforce such measures. "We can do things only step by step," he says. "It's not easy."

Sokhan

won't be pressed into further explanation, but Rozendaal is frank about the obstacles to progress. Corruption is endemic in Cambodia, he says, and the distributors of counterfeit drugs are protected by powerful figures in the government. Now working as a consultant in Indonesia, Rozendaal is free to speak out. But when he raised the alarm about fake drugs in Cambodia while running the European Union's malaria project, Rozendaal felt his job was threatened as spurious complaints about his competency reached his bosses in Brussels. For local officials, the stakes could be even higher. "You really take a big personal risk," says Rozendaal.

In some neighbouring countries, law enforcement is more rigorous. In Laos, for example, some pharmacies have been closed down. And as an example of what's possible, experts point to Nigeria, where one influential woman has made a courageous and effective stand against pharmaceutical counterfeiting (see 'In the line of fire', page 134).

But counterfeit drugs will continue to plague southeast Asia until the problem is defeated at its source. That means targeting the factories that are producing the fakes, which are mostly thought to be in China. This, in turn, will require pharmaceutical firms and governments to exert pressure on the Chinese authorities. Some progress is being made, says Joe Damond, deputy vice-president for international affairs with PhRMA, the trade body that represents the US drug industry. Regulations designed to combat drug counterfeiting have recently been strengthened, he says.

The acid test will be how these are implemented. So far, Chinese drug counterfeiters have mostly been handed modest fines that they can write off as a cost of doing business. As a result, some organized criminals are believed to have shifted their activities from narcotics, where sentences are severe and include the death penalty, to fake pharmaceuticals. Those who mourn Veasna's death can confirm that this new trade is equally murderous. And they are still waiting for his killers to be brought to justice.

Peter Aldhous is Nature's chief news & features editor.

1. Rozendaal, J. A. *Lancet* 357, 890 (2001).
2. Newton, P. et al. *Lancet* 357, 1948–1950 (2001).
3. Newton, P. N., White, N. J., Rozendaal, J. A. & Green, M. D. *Br. Med. J.* 324, 800–801 (2002).
4. Dondorp, A. M. et al. *Trop. Med. Int. Health* 9, 1241–1246 (2004).
5. *Study Report on Counterfeit and Substandard Drugs in Cambodia* (Cambodian Ministry of Health, Phnom Penh, 2001).
6. *Second Study Report on Counterfeit and Substandard Drugs in Cambodia* (Cambodian Ministry of Health, Phnom Penh, 2004).
7. Jahnke, R. W. O. *Pharm. Ind.* 66, 1187–1193 (2004).
8. Green, M. D., Mount, D. L., Wirtz, R. A. & White, N. J. *J. Pharm. Biomed. Anal.* 24, 65–70 (2000).

HIV drug remains unproven without placebo trial

Ethical concerns over use of a placebo weaken evidence for the benefits of nevirapine.

Sir—While raising concerns about “standards of record keeping” in the HIVNET 012 trial in Uganda, in your News story “Activists and researchers rally behind AIDS drug for mothers” (*Nature* **432**, 935; 2004), you overlook a greater flaw. None of the available evidence for nevirapine comes from a trial in which it was tested against a placebo. Yet, as the study’s senior author has said (see www.hopkinsmedicine.org/hmn/S01/feature.html), a placebo is the only way a scientist can assess a drug’s effectiveness with scientific certainty.

The HIVNET 012 trial abandoned its placebo group in early 1998 after only 19 of the 645 mothers randomized had been

treated, under pressure of complaints that the use of a placebo was unethical.

The HIV transmission rate reported for nevirapine in the HIVNET 012 study was 13.1%. However, without antiretroviral treatments, mother-to-child transmission rates of HIV vary from 12% to 48%. The HIVNET 012 outcome is higher than the 12% transmission rate reported in a prospective study of 561 African women given no antiretroviral treatment (J. Ladner *et al.* *J. Acquir. Immun. Def. Syndr. Hum. Retrovirol.* **18**, 293–298; 1998).

There are also reports of placebo-group transmission rates that vary within the same hospital and between hospitals, as

well as during different time periods of the same study. One study reported a lower transmission rate in the placebo group than with no treatment.

On what basis can it be claimed that “there’s nothing that has in any way invalidated the conclusion that single-dose nevirapine is effective for reducing mother-to-child transmission”? Without supporting evidence from a placebo-controlled randomized trial, such statements seem unwarranted.

Valendar F. Turner

Department of Health,
189 Royal Street, East Perth,
Western Australia 6004, Australia

Concern over deep-sea reefs is widespread

Sir—As a deep-sea biologist I was surprised to read the letter by Kjartan Hoydal from the North-East Atlantic Fisheries Commission (*Nature* **433**, 105; 2005).

Contrary to Hoydal’s contention that the News Feature “Sink or swim” (*Nature* **432**, 12–14; 2004) reflects a campaign by a few scientists and non-governmental organizations, the scientific evidence that fisheries are a threat to newly discovered deep-sea habitats is well documented and has been acknowledged by several governments.

The concerns raised in your News Feature are shared by many deep-sea biologists: 1,136 scientists from 69 countries signed a recent petition to the United Nations on this issue. Many governments and international bodies also believe that bottom trawling and other fishing practices are causing significant damage to deep-sea habitats.

Deep-sea coral reefs have been known for 200 years or more, but our understanding of their potential contribution to marine biodiversity is more recent, owing to new technologies that allow direct imaging of the deep-sea floor. This technology has also provided graphic evidence of destruction of these habitats by bottom trawling.

For example, it is estimated that off Norway up to 50% of reefs, formed by the coral *Lophelia pertusa*, have already been affected by fishing. In the Northern Rockall Trough, northwest of Scotland, evidence of damage to coral has been seen both in video images and in acoustic pictures of the seabed.

The United Nations Environment Programme (UNEP) recently published a report by several world experts on deep-sea coral reefs: *Cold-water Coral Reefs: Out of Sight, No Longer out of Mind* (UNEP-World Conservation Monitoring Centre Biodiversity Series No. 22), available at www.unep-wcmc.org. The report states: “Undoubtedly, the greatest and most irreversible damage is due to the increasing intensity of deep-water trawling that relies on the deployment of heavy gear which ‘steamrollers’ over the sea floor.”

This report was endorsed by environment ministers from several European nations. Because of quantitative evidence of trawling damage, the Norwegian government has banned bottom trawling from areas in which *L. pertusa* reefs occur. Measures have also been taken to protect *L. pertusa* in the northern Rockall Trough (European Commission Regulation no. 1475/2003).

The impacts of deep-sea trawling on coral reefs have been well documented in other parts of the world, including seas off Nova Scotia, the southern United States southern Australia.

I welcome the North-East Atlantic Fisheries Commission’s recent move to protect some deep-sea habitats in the North Atlantic. However, the fishing industry has little awareness of the environmental damage caused by trawling and other fishing practices.

Sustainable management of the deep-ocean environment will require a substantial and joint effort by marine scientists, fisheries managers and the fishermen themselves.

Alex David Rogers

Natural Environment Research Council,
British Antarctic Survey, High Cross,
Madingley Road, Cambridge CB3 0ET, UK

Brown knew particles were smaller than pollen

Sir—In your Year of Physics supplement (*Nature* **433**, 213–257; 2005), several authors repeat the mistaken idea that the botanist Robert Brown observed the motion that now carries his name while watching the irregular motion of pollen grains in water. The microscopic particles involved in the characteristic jiggling dance Brown described were much smaller clay particles. I have regularly studied pollen grains in water suspension under a microscope without ever observing brownian motion.

From the title of Brown’s 1828 paper “A Brief Account of Microscopical Observations ... on the Particles contained in the Pollen of Plants...”, it is clear that he knew he was looking at smaller particles (which he estimated at about 1/500 of an inch in diameter) than the pollen grains.

Having observed ‘vivid motion’ in these particles, he next wondered if they were alive, as they had come from a living plant. So he looked at particles from pollen collected from old herbarium sheets (and so presumably dead) but also found the motion. He then looked at powdered fossil plant material and finally inanimate material, which all showed similar motion.

Brown’s observations convinced him that life was not necessary for the movement of these microscopic particles. Brown was not the first to observe the motion that now carries his name and that Einstein famously explained. However, he was convinced his was the first really detailed study of the phenomenon and he clearly hoped for priority for his description.

David M. Wilkinson

Biological and Earth Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

A part of but apart from politics

Can scientists advise policy-makers without compromising their objectivity?

Nature's Experts: Science, Politics, and the Environment

by Stephen Bocking
Rutgers University Press: 2004. 304 pp.
\$60, £44.50 (hbk); \$24.95, £17.50 (pbk)

Roger Pielke Jr

In this excellent book on environmental science and politics, Stephen Bocking grapples with a problem that he characterizes as a riddle: "How can science be part of the political process yet separate?" Or more specifically: "How can we ensure that scientific research provides the information we need to pursue our environmental values and priorities (whether these relate to exploitation or to protection) without science itself becoming subject to the conflicts and controversies of environmental politics?"

For decades, the riddle posed by Bocking was answered through a widely shared conceptual model about the role of science in society, presented most influentially in Vannevar Bush's 1945 report to government, *Science: The Endless Frontier*. The policy advocated by Bush depended on a distinction between basic and applied research, with basic research contributing to a reservoir of knowledge that could be tapped to solve problems and exploit opportunities. Basic research was characterized simply as the quest for knowledge; it was pure. The elegance of the plan is that basic research was also fundamental to societal progress, and hence was part of the political process. Science was both separate from, and yet a part of, politics and decision making. This blueprint created momentum that was ultimately expressed in the creation of the US National Science Foundation. But the Bush report is remembered chiefly for its expression of the axiology and culture of science — it proposed that science should be led by scientists, not politicians — rather than for its ideas for creating institutions.

Bush's science policy has always had its critics, but only in the past decade or so have their criticisms been accompanied by shifts in science policy. Scholars of science and politics have begun to characterize conceptual models for science policy that move away from his model. An example is provided by Helga Nowotny, chairwoman of the European Research Advisory Board of the European Commission, and her colleagues. She says the old paradigm of scientific discovery "characterized by the hegemony of disciplinary science, with its strong sense of an internal hierarchy between the disciplines and driven by the autonomy of scientists and their host institutions, the universities" is



Valentine's message: passions can run high when scientists get involved in the political process.

being superseded, but not replaced, by a new paradigm of knowledge production. The new paradigm is "socially distributed, application-oriented, trans-disciplinary and subject to multiple accountabilities".

Another characterization of the move away from the old Bush model is the late Princeton historian Donald Stokes's 1997 model *Pasteur's Quadrant*, which reflects "use-inspired basic research" — a concept that is oxymoronic in the context of the Bush model. This sort of thinking has led, for example, to the adoption of the National Science Foundation's second review criterion, which focuses the evaluation of research proposals on societal impacts of research, and to ever more attention to scientific assessments and advisory bodies as important institutions in the policy process.

Scholars of science, technology and society have developed a robust body of knowledge that shows the flaws in the Bush science policy, and some parts of the community have experimented with alternatives. Even so, the Bush model is still widely embraced by many scientists and policy-makers. For instance, in 2004 the US National Research Council issued a report on scientific advisory committees, recommending that political considerations should not play a role in the process of deciding whom to appoint to policy panels. Yet its own membership, drawn from former government officials appointed by past presidents, reflected a perfect political balance of those appointed by Republicans and those appointed by

Democrats. Similarly, the Intergovernmental Panel on Climate Change has the temerity to claim that it is "policy neutral", yet its website trumpets its success in advocating the adoption of the Kyoto Protocol to the Framework Convention on Climate Change. As science policy has changed, these actions show signs of schizoid behaviour — the result of efforts to keep science both part of and separate from politics at a time of fundamental change in science policy.

In the context of this change in the role of science in society, Bocking clearly explains that the authority of science is ephemeral. He quotes Dorothy Nelkin: "As scientists debate the various sides of politicized issues, their involvement undermines assumptions about the objectivity of science, and these are precisely the assumptions that have given experts their power as the neutral arbiter of truth." By participating in political debates as advocates for special interests, scientists are reducing their claim to authority. Science becomes "politics by other means".

Two accessible chapters provide an overview of literature from the fields of science and technology studies, and policy. They cogently present many of the complexities and contradictions of science in policy and politics. Bocking presents well-informed discussions of three cases: natural resources management, global environmental change, and chemicals in the environment. Each chapter is well written and argued but is chock full of detail and allusion that may make them most meaningful to those

already familiar with each case.

Concepts such as 'credibility', 'influence' and 'deliberative democracy' are the focus of the book's closing chapters. Although they are important concepts, these chapters would have been more valuable if they had been discussed with the same analytical breadth and empirical depth of the book's first six chapters. But the same critique can be made of much of the literature on science and society: strong on diagnosis, less strong on prescription. It shows that scholars of science, policy and politics are just like the experts they study — they have more work to do in practising what they have learned about knowledge and action in the changing context of science policy.

Overall, this is an excellent book, worth reading by anyone interested in science, politics and the environment. But it is likely to be of particular value to environmental scientists who want to understand how and why the role of science in society is changing. ■

Roger Pielke Jr is at the Center for Science and Technology Policy Research, University of Colorado, Boulder, Colorado 80309-0488, USA.

Marketing Marie

Obsessive Genius: The Inner World of Marie Curie

by Barbara Goldsmith

Weidenfeld & Nicolson/W. W. Norton: 2004. 320 pp. £14.99/\$23.95

Susan Lindee

The life of Marie Curie was better than fiction. She was brilliant, driven and difficult; a scientist of the first rank; a mother; a young widow; a scorned woman in love with a married man; and a shrewd marketer of both herself and her discoveries. She won two Nobel prizes — for her work on radioactivity and the discovery of radium and polonium. She died aged 67 as a result of her cavalier approach to laboratory exposure to radioactive materials. Her appearance was also striking, both as a pale, beautiful young woman and as a stern, intense matriarch.

It is perhaps unsurprising that her life story has been told and retold in hundreds of biographies since her death in 1934. Beginning with her daughter Eve's affecting 1937 study, biographers have sought to illuminate her tragedies, intellectual style, determination and achievements, in studies aimed at scientists, children and the general public, published in English, French and many other languages. The flood shows no signs of abating: since 1995 there have been 36 English-language biographies entitled, roughly, *Marie Curie*. These include Susan Quinn's *Marie Curie: A Life* (Simon & Schuster, 1995), a fast-paced and well documented portrait of the Curies and their world, and *Marie Curie: A*



More than a woman: Marie Curie believed she had overcome the physical constraints of the body.

Biography by historian of science Marilyn Bailey Ogilvie (Greenwood, 2004).

It would be reasonable to wonder why anyone would want to write another biography of Marie Curie. Her personal papers, still somewhat radioactive, have been accessible to researchers for more than a decade, and the details of her life have been well known in outline for 60 years. What more could there possibly be to say?

Yet I must acknowledge that Barbara Goldsmith has managed to say some interesting things, and they are not the result of intense data-mining. Rather, she contributes a slow, methodical curiosity about matters that other authors have brushed past.

Goldsmith turns her attention to Curie's shifting ways of writing about her husband Pierre in the diary written after he was killed in a traffic accident, and to the mixed messages in Pierre and Marie's discussions of the commercial applications of their work. She addresses Marie's decision to involve her 17-year-old daughter Irene in the dangerous and gruesome war work of X-raying stricken soldiers, and she ponders Eve's estrangement from her mother. The Curies' interest in spiritualism, their attendance at séances and

their involvement with the Society for Psychological Research are explored in the context of Marie's reactions to Pierre's death. The public scandal over her affair with the physicist Paul Langevin is considered primarily in terms of its depressive effect on Marie, rather than in sociological terms, which could illuminate gender relations in early twentieth-century France. Goldsmith then considers the Curies' incautious handling of radium, which she attributes to their love of their own discovery. A final chapter outlines the family's enduring legacy and continuing scientific achievement. In all these discussions, Goldsmith makes good on her promise to excavate an "inner world".

As Goldsmith acknowledges, Marie Curie invented her own life story in the 'autobiographical notes' that accompanied her 1929 study of Pierre after his death. That life story, which has shaped virtually every biography of her since, emphasized the irrelevance of the physical body to scientific work. She described her paltry diet, her cold garret room and her poverty, using such details to highlight the legitimacy of her apprenticeship to science. Curie said she was strong enough to overcome the constraints of the

Science in culture

The Pied Piper of Düsseldorf

The artist Joseph Beuys tried to lead his followers into a promised land of transformative imagination.

Martin Kemp

Shaman or charlatan? Genius or joker? The German artist Joseph Beuys polarizes opinions. Sculptor, assembler of things, installation maker, performance artist, teacher and polemicist, Beuys became a kind of Pied Piper for generations of young artists in the final three decades of the twentieth century. Dressed in a sleeveless flak jacket and white shirt, with a wide-brimmed felt hat shading dark-rimmed eyes sunken in gaunt features, he exercised a mesmeric spell. He aspired to lead his followers into a promised land of new human potential and transformative imagination. Modest he was not.

He was determined to rewrite both history and the future. He was fascinated by Leonardo da Vinci, like so many others who have sought to transform our creativity, individually and collectively, by reuniting science and art. In the mid-1970s Beuys created his own set of drawn pages to emulate Leonardo's rediscovered codices in the National Library in Madrid. Leonardo is seen as marking a turning point in the history of Western thought. His creations were still imbued with the spiritual wholeness of medieval thought but showed clear signs of the positivist and materialist character that was to dominate subsequent science and technology and fragment our modern consciousness.

Beuys distrusted, or even hated, the analytical 'coldness' of modern science and the materialist mechanisms of our technological society. Such feelings are a recurrent theme in artists' commentaries. Benjamin Robert Haydon, an English painter of the early nineteenth century, declared that Isaac Newton had "destroyed the poetry of the rainbow, by reducing it to its prismatic colours". It is easy to dismiss such views as ignorant and puerile, but Beuys was familiar with science, having had a passion for natural history as a youth and flirting with medicine before training as an artist.

His imagery draws heavily on science and technology, particularly in the diagrammatic organization of philosophical, scientific, economic, political



and social concepts on the blackboards that featured in so many of his performances. His vitrines (works in glass cases), such as *Double Objects*, speak the vocabulary of display in science museums, and regularly exploit substances and objects that involve what might be called a 'humanized technology'. Within the glass case the paired objects, reverentially arrayed, play on recurrent themes in his work. There are typical echoes of the mythologized story of his rescue as a crashed fighter pilot on the Crimean front in 1944. His broken and freezing body was, legend has it, coated in fat, wrapped in felt and revived by Tartars who discovered his wrecked plane.

The battered batteries, blocks of peat, coal, butter and soap speak of energizing, vivifying, warming and cleansing. The telephone, made primitively from double tins, and the Siberian Symphony gramophone records evoke poignant sounds from past times. The enamelled bowls, glass bottles and paired X-ray images together convey medical associations. In Beuys' personal language of symbols, the X-rays and the brown crosses painted on the telephone represent

beneficial qualities, and the doubling of the objects signifies human conjoining and completeness.

What Beuys was striving to assert on the broadest front of social communication was the centrality and unifying power of the warmth and wholeness of the human spirit in all the diverse fields of our activity. Almost 20 years after his death, his agenda can be seen to belong to the long history of utopian visions that are more seductive than realizable. What does remain, and continues to exercise its power, is an uncanny evocation of his personal presence, his tale of near-death and his charismatic life.

His surviving creations, many of which can be seen at Tate Modern in London until 2 May, transform ordinary items and mundane substances into composite entities. They speak of human values that science must address and embody if it is to be integrated and comprehended in our twenty-first century society.

Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK, and co-director of Wallace Kemp Artakt.

♦ www.tate.org.uk/modern/exhibitions/beuys

body, its need for food and warmth. By implication, she could also overcome those constraints that made her own, female body an almost insurmountable barrier to intellectual achievement. Her descriptions of a self-sacrificing life of the mind had an enduring popular appeal, and biographers writing for the children's market have emphasized Curie's indifference to 'feminine' things of the body (food, clothing, beauty) and her engagement with 'masculine' things of the mind (science, truth, evidence and power).

Goldsmith, an experienced biographer, notes that she was herself seduced as a young

girl by the Curie legend. Like myself, and so many others, she was drawn to the image of the brilliant, tragic, woman scientist, particularly as played by Greer Garson in the 1943 film *Madame Curie*. In one memorable scene, the newly introduced Marie and Pierre begin to discuss their scientific work, but just as the first few technical words are exchanged, the scene suddenly leaps forward in time to the close of their discussion and the last few technical words. The intellectual life that bound Marie and Pierre together apparently had no particular bearing on the love story. Goldsmith's biography is similarly unconcerned

with Curie's scientific contributions. Quinn's 1995 account presents Curie's scientific world in much more detail, but a full-length scientific biography of Curie has yet to appear.

Goldsmith has the sense to refrain from grandiose and dramatic claims, and her tone throughout is quizzical, calm and perceptive. The book provides not new information but a thoughtful perspective on the life of one of the most important scientists of the twentieth century. The author is careful in her extrapolations from available records, and does not assume that she knows anyone's inner feelings unless they have been expressed

in writing. With its clear and accessible style, the book could be shared with young readers, who might be less susceptible than earlier generations to narratives of romantic self-sacrifice, and more intrigued by the psychological portrait of a complicated and accomplished woman scientist. ■

Susan Lindee is in the Department of History and Sociology of Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6304, USA.

Rivalry and revenge

Costantinopoli 1786: la congiura e la beffa (Constantinople 1786: The Conspiracy and the Hoax)

by Paolo Mazzarello

Bollati Boringhieri: 2004. 327 pp. €24.

In Italian.

<http://www.bollatiboringhieri.it/>

Nicola Nosengo

The second half of the eighteenth century was a time of spectacular advances in the life sciences. Fundamental problems such as the generation of life were addressed for the first time using modern experimental tools. But these issues were the source of great controversy, and also great rivalries among biologists — or philosophers, as they still preferred to call themselves.

At a time when many scientists were still convinced that life can be generated spontaneously from decomposition, the Italian Lazzaro Spallanzani was the first to demonstrate the necessity of sperm for reproduction. He was the first to perform artificial insemination — a challenging experiment at the time, especially as he was a Catholic priest.

Spallanzani was a strong opponent of Carolus Linnaeus, who in 1735 had published the first complete classification of living species, mostly based on external traits. Spallanzani, however, was convinced that naturalists should study every aspect of living forms, including their behaviour and environment, rather than simply force them into groups according to what they look like.

Such a man could easily make enemies. In Pavia, where he taught natural history at the university and was responsible for the museum, Spallanzani was surrounded by envy, particularly from Giovanni Antonio Scopoli, a chemist and botanist, and devoted follower of Linnaeus. The rivalry between Spallanzani and Scopoli resulted in a conspiracy worthy of a Shakespearean tragedy. The tale is told in this book by Paolo Mazzarello, a professor of the history of medicine at the very same university in Pavia.

In 1785, Spallanzani set sail for Constantinople, taking advantage of Venice's new diplomatic links with the Ottoman capital. He had taken a year off from teaching in Pavia, but this was no pleasure cruise — it



Lazzaro Spallanzani poured scorn on rivals whose experiments failed to meet his own high standards.

was a genuine scientific mission. Spallanzani left Pavia equipped with scientific instruments, such as barometers, thermometers, lenses and a microscope. He spent most of his time taking measurements and collecting samples, studying everything from living species to climate and geology.

In Spallanzani's absence, his rivals conspired against him. Under the instructions of Scopoli, Serafino Volta, Spallanzani's assistant in the museum, visited his family home in Emilia, asking to see his private museum. He then accused Spallanzani of stealing objects from the Pavia museum and putting them in his own collection. The news spread throughout Europe and reached Spallanzani, by now in Vienna on his way home.

On his arrival in Pavia, Spallanzani was able to prove his innocence and restore his reputation. His rivals were severely reprimanded by the authorities in Milan.

But this was not enough for Spallanzani, who took a supplementary revenge when Scopoli, in 1785, announced his discovery of a new species of worm: a human parasite unlike any other previously known. The parasite ostensibly came from a woman from northern Italy, and its scientific description brought Scopoli fame. He even dedicated it to the botanist Joseph Banks, then president of the Royal Society in London. But in 1787 the worm turned out to be a hoax, constructed by a farmer for a joke from the oesophagus and windpipe of a chicken. Spallanzani ensured that Scopoli's humiliation was widely

known. Using a pen name, he wrote a pamphlet, full of scorn and cruel irony, condemning Scopoli's ability as a scientist.

Scopoli, he wrote, wanted to study nature inside "dead museums", only hoping to be lucky enough to discover and classify new animals and put them on the shelves. But nature, in Spallanzani's view, had to be studied in action, observing animals' behaviour, comparing them in terms of the structures and functions of their organs. Had Scopoli paid more attention to the internal structure of the worm, had he stopped to wonder how it could actually live, he would probably have realized his mistake and avoided his humiliation, argued Spallanzani in his pamphlet.

Mazzarello tells the whole story with detailed descriptions of landscapes and the people Spallanzani met during his journey, based on his diaries. He describes Spallanzani's "lust for knowledge", his being as "fascinated and seduced by Nature as his contemporary Giacomo Casanova was by women".

But what is really compelling is that through this plot we can observe the development of experimental methods in biology. It is mostly over methods that Spallanzani and his rivals clashed. The worst that Spallanzani could throw at his rival when he took his revenge was to call him a poor experimenter. Personal authority was no longer enough in science unless it was supported by reliable and replicable methods. ■

Nicola Nosengo is on the editorial staff of the Italian science magazine Galileo, based in Rome.

Schrödinger's mousetrap

Part 8: The outcast.

Jack Cohen

Lister sat quietly, waiting for Jirong Feng's interview. He felt somewhat awed by what he had seen of the oriental; he looked to Lister's Western eyes about 22, but he was probably older. Everyone had said how bright he was, with the apparently erratic approach of the truly original thinker. Could he have set up this convoluted crime? Feng came quietly into the room and stood by the interview chair. He turned it around in a quick, graceful movement and sat with his arms on the chair-back, a definite twinkle in his eyes as they met Lister's. Lister smiled, too.

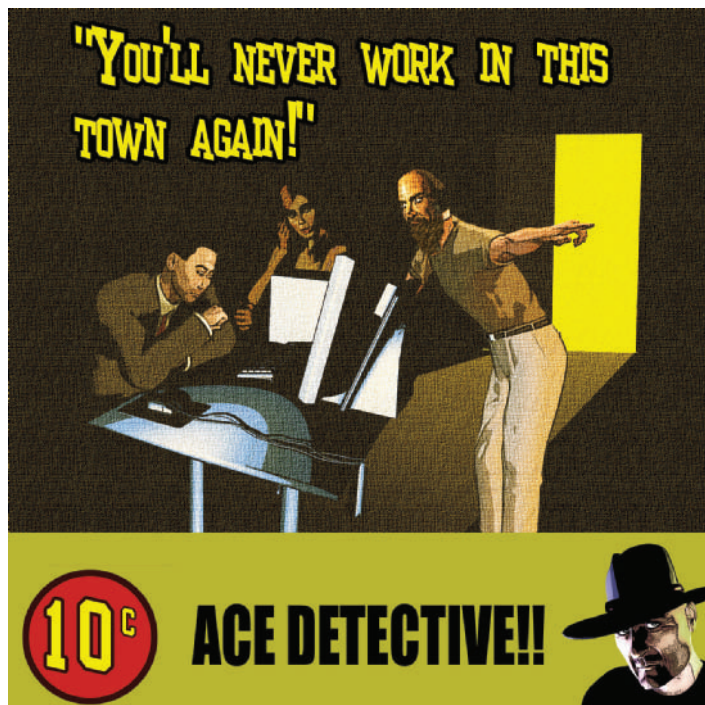
"I had nothing to do with it," Feng said, "at least directly. And I feel that it would be disloyal to some of my friends, my benefactors, simply to spill everything I know about them to you. But I'll respond to your questions as honestly as I can, I want to see this cleared up like we all do. Or all but one, I suppose." And he twinkled again. "Rufus I won't miss. But I've not been carrying a grudge. Everything has worked out so well for me now I've left that gang."

"The University of Wentbridge gave you a black mark, though, didn't they?" Lister asked. "I'd have thought you would feel that Professor Jaeger could have supported you better..."

"Sure he could, but that was not his style: he had to be seen to be doing the correct thing. He often fell on to his back in an attempt not to lean too far forward, as we say in Korea. He sacrificed me to his reputation; but he would have justified that to himself with consideration of all the future postdocs he could help with his reputation intact."

"Is it relevant to my questioning here, what went wrong?" Lister asked. "Is there someone, perhaps Professor Jaeger himself, who knows something about those experimental results that you want to keep secret?"

"No, absolutely not," said Feng quietly. "I'll give you my notebooks of the time, and you'll see what happened. Some measurements were confused as they accumulated in the spectrometer; I took them for mine but they were someone else's. I made



up a great theory assuming they were mine. It's old history now, it was a silly business, all my fault really. But unlike some bosses, unlike Petra — that's Professor Pruszczyński, my present boss — Rufus never forgave anything. Petra says she forgives 10%, it makes for an easier life, and I think I'll adopt that practice. My PhD work has really taken off, you know... much better now."

Lister leaned forward. "Can I ask you why it is that your boss is so spiky — was so spiky — towards Professor Jaeger? Was it just professional antipathy, or was there a deeper reason?"

"Come on," said Feng, "there can't be a deeper reason for Petra. It's not 'just' professionalism, it's scientific integrity; that's the most important thing in the world for my boss..."

"So very important... Mmmm. Now it was Ludmilla Shlomiuka who recommended you to your new boss some time after you were sent away by Jaeger, wasn't it?"

"Yes, it was, and I want to tell you how grateful I am to her. She has such big troubles in her life, but she likes everybody, is helpful to everybody — she

wrote to Petra recommending me, but she could have lost her job if Rufus had found out."

"And have you kept up your acquaintance with Jaeger's group?" Lister asked.

Feng stared at him for a moment, then

said: "I don't have any silly loyalty to Rufus's lot as a research group. Veronique Dubois, who regularly visits my boss, has been buying me drinks, if you know what I mean, and I haven't been holding back any of the technical stuff I learned as Rufus's student. But most of it's public anyway, they didn't invent much there. Working for Petra is much more interesting. There's a dozen tricks that we've developed that Rufus would've given his eye-teeth for, as you people say."

"But yes, I am occasionally in contact with Ludmilla... I'm very grateful to her and I try to be as helpful as I can now and then as she is going through a difficult time with her husband Dmitri."

"Oh, yes, you wouldn't know. He was one of the early implant cases, fetal nervous

tissue implanted to cure his Parkinson's, they do it properly now, but then... Anyway, his moods are all over the place, he has gone through very violent spells, had to be taken in to the hospital by the police sometimes. She's fairly successful at holding him stable now, but it hasn't been easy. Poor Ludmilla!"

Feng was beginning to tell Lister about his present research with negadex and other materials, and the detective's mind wandered for a moment as he lost touch with the esoteric physics. Feng suddenly realized he'd lost his audience. "Sorry about that," he said. "I get carried away. Is there anything else I can tell you about?"

"Where were you during the coffee break?"

"I had met a couple of women from my old university department in Korea, come over here to do research. I'm afraid that my mind was not on the physics at all, but on how I could get to see more of them. We were interrupted by Wilfred de Bruijn, who shouted at me a bit about my questions to him yesterday after his talk. Then Nigel Lorimer came and took him away. My women friends were not impressed. Do you want their address? They have nothing to do with this affair, I promise you."

To be continued...

Jack Cohen is an author and reproductive biologist based in Newent, Gloucestershire, UK.

Who do you think killed Rufus Jaeger? Catch up on all the evidence and vote for your suspect at www.nature.com/news/mousetrap

Eggs alone

Human parthenotes: an ethical source of stem cells for therapies?

Ann A. Kiessling

Eggs are enormous cells in a state of arrested expectancy. Before the advent of assisted reproductive technologies, human eggs played out their fate in the privacy of the ovary, fallopian tubes and uterus. Women are born with around one million eggs, but by 55 years of age, all these are gone. The loss appears to be relatively uniform throughout life — each year approximately 20,000 eggs die in the ovary, whereas just 12 mature and ovulate. Once released into the fallopian tube, an egg that encounters sperm is fertilized and begins the rapid cycles of cell division that are characteristic of the normal embryo. It completes its journey to the uterus, where it develops the small group of stem cells with the potential to differentiate into all tissue types, a characteristic known as 'pluripotency'. Eggs not fertilized by sperm were simply presumed to disintegrate.

But eggs can activate on their own without sperm, a process known as 'parthenogenesis', from the Greek for 'virgin'. Although some insect species produce workers by parthenogenesis, mammalian parthenotes fail to develop into offspring. Mouse and rabbit parthenotes appear morphologically indistinguishable from fertilized eggs for the first few days of development, including the formation of the small group of pluripotent stem cells. But these parthenotes are not viable and fail to develop into offspring. Two types of growths — dermoid cysts of the ovary and teratomas — suggest that human eggs can also undergo parthenogenesis. Dermoid cysts arise if the ovarian sack fails to rupture; the retained egg activates cell division on its own and gives rise to a small cluster of cells that undergo limited differentiation. Teratomas are benign growths that are not restricted to the ovary and exhibit a variety of cell types, including skin, muscle, bone and occasionally hair and rudimentary teeth.

Long regarded as curious, unimportant quirks of nature, teratomas and ovarian dermoids provide evidence that human parthenogenesis does occur and that human parthenotes spontaneously give rise to at least a few pluripotent stem cells.

Embryonic stem cells from mice have been used as experimental tools for more than two decades, but widespread recognition of the therapeutic potential of pluripotent human stem cells is more recent. Stem-cell therapy offers hope for diseases that result from a lack of a tissue reservoir of stem cells to replace defective and dying cells, such as heart failure, diabetes, Parkinson's



A human parthenote 12 hours after activation.

disease, spinal-cord injury, AIDS and perhaps Alzheimer's disease. But the use of human embryos as a source of stem cells has proved morally and ethically contentious, and led to bans on human embryo research in several countries, including the United States. Could human parthenotes provide an alternative to the use of fertilized eggs?

One caveat to the potential therapeutic value of parthenote stem cells centres on the very reason that mammalian parthenotes do not develop to offspring. Genes inherited from the mother are expressed differently during embryonic development from genes that are inherited from the father. This trait is termed genetic 'imprinting', and it is not known with certainty that stem cells carrying only maternally imprinted genes will behave as robustly as embryonic stem cells. However, it is encouraging that the one line of monkey parthenote stem cells that has been reported (Cyno-1) seems to have the same potential as embryonic stem cells to continually divide and to develop into all cell types.

At present, the only source of human eggs is the human ovary. Is it possible to rescue some of the thousands of doomed eggs for use as parthenote stem cells? Or do these considerations simply replace the moral dilemma of using human embryos for obtaining stem cells with the moral dilemma of collecting eggs from the ovaries of women for therapeutic instead of reproductive purposes? The derivation of

stem-cell lines from parthenotes depends on the availability of human eggs and on developing the technology to activate them in the laboratory and promote cell division to stem cells. Proof of the principle was provided by two reports, but no stem-cell lines were derived.

For women with serious diseases, such as Type 1 diabetes or spinal-cord injury, using their eggs to generate stem cells for their own potential therapeutic use requires little justification. The process could be likened to self-donation of bone-marrow stem cells, or blood before a surgical procedure. Because genetic information is exchanged between paired chromosomes during egg maturation, parthenote stem cells may not be genetically identical to the woman, but will undoubtedly be a closer tissue match than stem cells from human embryos.

Moreover, it is theoretically possible that such parthenote stem-cell lines could also be made available to other tissue-matched individuals, including men or older women. In this way, parthenote stem-cell banks could replace the need for the ethically controversial embryonic stem-cell banks.

A recent elegant experiment has added to the confusion about the 'embryo' status of parthenotes. Genetically engineered mouse egg chromosomes were shown to behave like mouse sperm chromosomes when transferred into an unfertilized mouse egg. The development of the resulting reconstructed egg was titled 'parthenogenesis', although it differed markedly from canonical parthenogenesis, which relies solely on the egg's own chromosomes.

Manipulating human eggs to give rise to stem cells for therapies instead of embryos for reproduction requires society to expand its understanding of the capabilities of eggs. New terms to accurately describe the new tasks will make it easier for policy makers to understand and establish appropriate guidelines for human egg research, including the promise of parthenotes. ■

Ann A. Kiessling is at the Harvard Institutes of Medicine, 4 Blackfan Circle, Boston, Massachusetts 02115, USA.

FURTHER READING

Kiessling, A. A. & Anderson S. C. *Human Embryonic Stem Cells* (Jones and Bartlett, Sudbury, 2003).

Kiessling, A. A. *Connecticut Law Rev.* **36**, 1051–1092 (2004).

Cibelli, J. B. *et al. Science* **295**, 819 (2002).

Kono, T. *et al. Nature* **428**, 860–864 (2004).

Fossils make waves

James W. Kirchner and Anne Weil

A 62-million-year cycle in biodiversity emerges from scrutiny of a marine-fossil database, but its causes remain mysterious. Thus, this discovery is likely to provoke a flurry of theoretical speculation.

For decades, palaeobiologists have been finding large-scale cycles and patterns in fossil records, and one might assume that there was nothing substantial left to be discovered. Not by a country mile! On page 208 of this issue¹, Rohde and Muller demonstrate a 62-million-year cycle in fossil biodiversity during the Phanerozoic, the past 500-plus million years of life on Earth. This 62-million-year wave is surprisingly strong and — so far — there's no good explanation for it.

In some fields of science, particular data sets achieve iconic status, focusing the attention of whole communities of researchers as they try to understand the patterns they contain. For the past two decades, many palaeobiologists have been exploring the marine-fossil databases created by the late Jack Sepkoski^{2,3}. Sepkoski spent years combing through the palaeontological literature, compiling the first and last known occurrences of every animal genus in the marine-fossil record, more than 30,000 in all. He used genera because, being the taxonomic level above species, they are revised less often by palaeontologists and their numbers are more manageable. And he dealt exclusively with marine animals, because their fossil record is longer and better preserved than that of their terrestrial counterparts. Sepkoski's labours produced a detailed history of biodiversity through time, which in turn has generated a kaleidoscope of analyses^{4–9} aimed at both detecting its patterns and explaining them.

Over time, these analyses have explored increasingly subtle features of the fossil record (Box 1), but Rohde and Muller's discovery shows that there are still broad patterns left to be found. They have taken Sepkoski's fossil diversity curve — a graph of the total number of marine-fossil genera at each point in the past 542 million years — and subtracted a smooth mathematical function that accounts for the broadest trends of diversity through time. What is left is a diversity curve that rises and falls somewhat irregularly, but roughly in step with a 62-million-year cycle. Rohde and Muller use spectral analysis to show that these ups and downs oscillate much more rhythmically than one would expect by chance.

The 62-million-year wave is too big to ignore, with biodiversity growing and shrinking by several hundred genera between the peaks and the troughs. The cycle is visually obvious, and some palaeobiologists will wonder why they haven't seen it before. In fact, a similar long wave in diversity was suggested almost 30 years ago in the pages of this journal¹⁰, but only with the most recent refinements to the geological timescale can the cycle in fossil diversity be shown to be too regular to have arisen at random.

Rohde and Muller show that the 62-million-year cycle is particularly strong when one considers only short-lived genera, which

they define as those that survive less than 45 million years. This seems to make perfect sense; longer-lived genera are thought to survive longer because they are more diverse and more widespread, and thus more resistant to disturbance. But it is also almost true by definition, because it would be hard to construct a wave from components that endure for more than half the wavelength, and that would therefore bridge the peaks and troughs.

Two other aspects of Rohde and Muller's discovery may also attract sceptical attention. First, the 62-million-year cycle is largely absent from the past 150 million years, which is precisely the part of the fossil record that is most accurately known. Second, many will have nagging suspicions that the cycle is somehow a statistical fluke, because some apparent patterns in the fossil record have evaporated when subjected to statistical scrutiny. To their credit, Rohde and Muller have tested the 62-million-year cycle in several different ways, and in each case it seems to be statistically significant.

If the 62-million-year cycle is real, it demands an explanation, and the search for one will be interesting. Rohde and Muller are physicists, and they examined 14 possible geophysical and astronomical drivers for the cycle. Several of these are at best highly speculative, including a companion star to the Sun and a hypothesized large planet for which no evidence exists. But in any case, all of these

"The 62-million-year wave is too big to ignore ... palaeobiologists will wonder why they haven't seen it before."

Box 1 Explorations of the fossil record

Palaeontologists use compilations of fossil taxa (which may be species, genera, families or even larger groupings) to study not only taxon richness through time, but also extinction and evolutionary diversification, and the rates at which these processes occur. As documented by Miller¹¹, the earliest curve of Phanerozoic species richness was constructed in 1860, and ever since then palaeontologists have striven for increasingly more complete and accurate data compilations.

In 1860, statistics was in its infancy. But by 1982, when Sepkoski published his database of fossil marine families¹², he did so with every intention of applying statistical tests. The best-known finding from this work was the hotly debated 26-million-year periodicity of extinctions. Further findings characterized the taxonomic composition of Earth's oceans through time, and evolutionary patterns common to different families. Sepkoski's database of fossil marine genera³ could be

parsed more finely; studies based on it have included several spectral analyses, many estimates of speciation rates, and calculations of how such rates might be influenced by the way in which time intervals are defined in geological timescales.

Moreover, many hypothesized causes of large-scale patterns depend on processes at smaller scales, and global databases are often unsuitable for tests requiring nuanced comparisons. Workers thus use more-detailed databases, or break down the

larger ones by geographical region, by time period, by taxon (snails vs clams, for example), by ecological role (predator vs prey; specialist vs generalist), or by environment (terrestrial vs marine; deep vs shallow water). Such partitioning enables researchers to study, for instance, spatial and temporal scaling; variation in the speed, taxonomic composition and geography of recovery from mass extinctions; and patterns of morphological diversification within large groups.

J.W.K. & A.W.

candidate driving mechanisms seem to be insufficient to cause the 62-million-year cycle.

As other disciplines enter the fray, the range of possible explanations will grow. Earth scientists and palaeontologists will point out that marine-fossil diversity depends on the diversity of marine habitats, and thus on the size and configuration of the continental shelves. They will therefore ask whether the 62-million-year cycle could potentially reflect changes in the continental margins through time, as sea level fluctuates and the continents rearrange themselves. Others will observe that this long wave in biodiversity is broadly consistent with the reported phase shift between fluctuations in the rate of extinction of existing organisms and the diversification of new ones⁷, and will search for a theory that unites these observations. Theoretical biologists will also note that global biodiversity is a tapestry that weaves itself, so the 62-million-year cycle in fossil diversity need not be generated by similar cycles in external driving factors. Instead, biodiversity could swing like a pendulum, with a rhythmic cycle that is governed by its own internal dynamics rather than by rhythmic external forcing.

But if the 62-million-year cycle is caused by a biological pendulum, it swings so slowly that it will be challenging to discern the underlying mechanisms. By any biological yardstick, 62 million years is a very long time; 62 million years ago last Tuesday, we mammals had only recently embarked on our striking morphological diversification following the mass extinction at the end of the Cretaceous. Clever modellers should have little difficulty creating biological models that exhibit very long oscillations in biodiversity. But the hard work will lie in showing that the premises behind these models are themselves accurate, or at least plausible.

It is often said that the best discoveries in science are those that raise more questions than they answer, and that is certainly the case here. Let the theorizing begin. ■

James W. Kirchner is in the Department of Earth and Planetary Science, University of California, Berkeley, California 94720-4767, USA.

e-mail: kirchner@seismo.berkeley.edu

Anne Weil is in the Department of Biological Anthropology and Anatomy, Duke University, Durham, North Carolina 27708-0383, USA.

e-mail: annweil@duke.edu

1. Rohde, R. A. & Muller, R. A. *Nature* **434**, 208–210 (2005).
2. Sepkoski, J. J. Jr *Paleobiology* **19**, 43–51 (1993).
3. Sepkoski, J. J. Jr in *A Compendium of Fossil Marine Animal Genera* (eds Jablonski, D. & Foote, M.) *Bull. Am. Paleontol.* No. 363 (Paleontological Research Institution, Ithaca, 2002).
4. Raup, D. M. & Sepkoski, J. J. Jr *Science* **231**, 833–836 (1986).
5. Miller, A. I. *Science* **281**, 1157–1160 (1998).
6. Jablonski, D. *Science* **284**, 2114–2116 (1999).
7. Kirchner, J. W. & Weil, A. *Nature* **404**, 177–180 (2000).
8. Peters, S. E. & Foote, M. *Nature* **416**, 420–424 (2002).
9. Miller, A. I. & Foote, M. *Science* **302**, 1030–1032 (2003).
10. Thomson, K. S. *Nature* **261**, 578–580 (1976).
11. Miller, A. I. in *Deep Time: Paleobiology's Perspective* (eds Erwin, D. H. & Wing, S. L.) 53–73 (Univ. Chicago Press, 2000).
12. Sepkoski, J. J. Jr *Milwaukee Public Mus. Contrib. Biol. Geol.* **51**, 1–125 (1982).

Astronomy

Stellar mass limited

Pavel Kroupa

Is there an upper limit to the mass of a star? The answer to this long-standing question seems to be yes — and it has important consequences for our understanding of the evolution of galaxies.

Stars form from interstellar gas and synthesize elements heavier than helium in their cores. Carbon, oxygen, silicon and iron, which are crucial for the existence of planets — and for life — are produced mostly by stars that are more massive than the Sun. To understand how galaxies evolve and are enriched in these elements, we need to know exactly how massive stars can be. Using the Hubble Space Telescope, Donald Figer (page 192 of this issue)¹ has analysed the stellar content of the Arches cluster, a highly populous, nearby star cluster. He finds that there are no stars heavier than about 150 solar masses, providing direct evidence for the idea that the spectrum of stellar masses has an upper limit.

The most abundant stars in the Universe are faint and cool red dwarfs^{2,3}. These have smaller masses than that of the Sun, with a lower limit of 0.072 solar masses⁴. Below this limit, the dwarf stars weigh so little that their central densities and temperatures are insufficient to stabilize them through thermonuclear reactions — they become 'brown dwarfs', and cool indefinitely. Most of the stellar mass of galaxies is locked up in long-lived dwarf stars, which return hardly any of their mass back to the interstellar gas from which they were born.

Stars that are heavier than the Sun rarely weigh more than 20 solar masses — for every 1,000 dwarf stars we find just one star weighing 20 solar masses, and more massive stars are rarer still². Stellar luminosity increases rapidly with mass, and whereas dwarf stars contribute very little to the luminosity of galaxies, massive stars live expensively, radiating some 10,000 times more energy per second than the Sun. Not surprisingly, they have a short lifespan: the Sun will live for about 10,000 million years, but stars heavier than about 20 solar masses live only a few million years.

Once the central region of a massive star has exhausted its primary fuel — hydrogen — it contracts, thereby increasing its central density and temperature further so that carbon and consecutively higher elements can be synthesized. Dwarf stars of solar mass or less never reach the conditions for synthesizing significant amounts of elements as heavy as oxygen or iron.

Eventually, when the core of a massive star has become sufficiently enriched in iron, it cannot be stabilized by the fusion of higher elements, and it collapses. An object then

forms that has a high density of neutrons — a neutron star — or under more extreme conditions a black hole may form. A supernova explosion ensues as the rest of the star crashes into the core, driving a detonation shock-front outwards and tearing most of the star apart. A single explosion can inject large amounts of heavy elements back into the galaxy and thus into the cycle of stellar birth and death. Attempts to understand the details of these processes lie at the forefront of astrophysical research, but theoretical calculations cannot convincingly reproduce the observed events. They often result in a whole star disappearing from our horizon by imploding to a black hole without a supernova explosion⁵ — at least in computer simulations.

It was not clear until now whether stars as massive as 300 or even 1,000 solar masses could exist. Theoretical modelling of such stars^{6,7} is difficult because of the extreme conditions of their interiors and their highly dynamical evolution. One theory⁸ states that stars more massive than about ten solar masses cannot form because the pressure exerted by radiation is so high that it prevents any further matter accreting from its natal cloud. The existence of heavier stars was therefore explained by proposing⁸ that they form from colliding and coagulating protostars. These weigh up to ten solar masses and are found in dense, central regions of young star clusters. Other theoretical work⁹, as well as observations¹⁰, suggest that accretion from massive circumstellar disks may in fact overcome the radiation pressure exerted by protostars heavier than ten solar masses.

To determine whether there is a stellar mass maximum, astronomers need to study rich populations of young stars such as are found in interacting galaxies (Fig. 1). But these distant populations cannot be resolved into individual stars, even with today's best telescopes. The Arches cluster studied by Figer¹ seems a suitable alternative for the task, as it is only a few million years old and is relatively near. It is rich in heavy elements, but lies very close to the centre of our Galaxy, making it difficult to observe. Yet in this populous cluster, stars more massive than about 150 solar masses ought to be present in sufficient numbers to be detected. However, they are entirely absent from Figer's observations and thus he concludes that they are unable to form.

Although Figer's result is statistically



Figure 1 Clash of giants — the colliding Antennae galaxies, observed with the Hubble Space Telescope¹². The patchwork of blue knots are star bursts, where tens of thousands of massive stars are forming. The existence of possibly hundreds of supermassive stars drives the highly violent events associated with star bursts. To fully understand such observations, we need to know the maximum possible mass of a star. This image conveys events similar to those that shaped our present-day galaxies shortly after the birth of the Universe.

highly significant, the uncertain age of the cluster leaves some room for arguing that it could be old enough for more massive stars to have already exploded as supernovae. However, the absence of hot, expanding gas bubbles does not support this scenario. In addition, observations in the similarly young and massive R136 cluster located in a neighbouring galaxy — the heavy-element-deficient Large Magellanic Cloud — support Figer's analysis, as here also no stars more massive than about 150 solar masses were found¹¹.

These results suggest that the upper limit for the mass of a star may be unrelated to the heavy-element content of the star-forming gas. This could in turn imply that radiation pressure is not the physical mechanism that limits how massive stars can become. Environments that are rich in heavy elements are dusty, and radiation pressure is particularly efficient at blocking the infall of dusty gas, so we might have expected the heavy-element-poor environment of the Large Magellanic Cloud to have allowed stars with higher mass to form through accretion.

Whereas the physics of faint, low-mass stars transcending into brown dwarfs is well understood⁴, there is no clear explanation for

why stellar masses should be limited to near 150 solar masses. And a nagging uncertainty remains: the details of supernova mechanisms are not fully understood, and it may be that stars more massive than 150 solar masses did exist, but have already imploded to black holes in the Arches and R136 clusters, leaving little trace except a hole in space-time accompanied by a brief burst of neutrinos and gravitational waves. ■

Pavel Kroupa is at the Sternwarte of the University of Bonn, Auf dem Hugel 71, D-53121 Bonn, Germany.

e-mail: pavel@astro.uni-bonn.de

1. Figer, D. F. *Nature* **434**, 192–194 (2005).
2. Reid, I. N., Gizis, J. E. & Hawley, S. L. *Astron. J.* **124**, 2721–2738 (2002).
3. Kroupa, P. *Science* **295**, 82–91 (2002).
4. Chabrier, G. & Baraffe, I. *Annu. Rev. Astron. Astrophys.* **38**, 337–377 (2000).
5. Janka, H. T. *Astron. Astrophys.* **368**, 527–560 (2001).
6. Hirschi, R., Meynet, G. & Maeder, A. *Astron. Astrophys.* **425**, 649–670 (2004).
7. Heger, A., Fryer, C. L., Woosley, S. E., Langer, N. & Hartmann, D. H. *Astron. Astrophys. J.* **591**, 288–300 (2003).
8. Bonnell, I. A., Bate, M. R. & Zinnecker, H. *Mon. Not. R. Astron. Soc.* **298**, 93–102 (1998).
9. Yorke, H. W. & Sonnhalter, C. *Astrophys. J.* **569**, 846–862 (2002).
10. Chini, R. *et al.* *Nature* **429**, 155–157 (2004).
11. Weidner, C. & Kroupa, P. *Mon. Not. R. Astron. Soc.* **348**, 187–191 (2004).
12. Whitmore, B. C. *et al.* *Astron. J.* **118**, 1551–1576 (1999).



100 YEARS AGO

Prof. A. H. R. Buller, writing from the University of Manitoba, describes some striking electrical effects due to the dryness of the atmosphere at Winnipeg... When the thermometer is low, ranging as it often does for a week or more at a time from 0° to –40° F., very little friction, such, for instance, as may be produced by walking along a carpet, causes a person to become charged with sufficient electricity to produce a visible and audible spark on touching an iron bedpost, the radiator, the gas-tap, or any other conductor. It is a favourite amusement of some children to take sparks from each other's noses after running about a carpeted room... Many ladies have considerable difficulty in combing their hair; for during the process it becomes so charged with electricity that it stands out in the most astonishing manner... It is quite easy to light the gas with a spark from the finger when matches are not handy by merely shuffling a few paces over the carpet and then holding a finger to the burner. From *Nature* 9 March 1905.

50 YEARS AGO

"Scientific Progress and Security Regulations." ... Dr. Hildebrand insists that positive achievement and progress, not the negative policy of restriction and security, provide the only firm basis of security... Security-screening programmes are a means to an end, not an end in themselves. Their role in defence policy is negative rather than positive. They may deprive a potential enemy, at any rate temporarily, of information about armed forces or the development of new weapons; but they create no new weapons themselves... What has now to be recognized is that, with the essential dependence to-day of military strength upon science, the security practices used in the past to safeguard military information are no longer fully valid. Scientific knowledge cannot be kept secret by such means. Progress in science is a cumulative process in which each scientist builds upon what is known, and national boundaries and security systems cannot contain this process of extending knowledge without so discouraging the spirit of inquiry that the state of learning and of technology as well as the rate of scientific progress are adversely affected. At best, among advanced nations, security measures can provide an advantage of time: there is no such thing as a permanent scientific secret. From *Nature* 12 March 1955.

Epigenetics

Surveillance team against cancer

George Klein

Variations in the control of a phenomenon known as parental imprinting influence the likelihood of tumour development. These new findings may tie in with an earlier concept of 'two-phase' carcinogenesis.

Cancer is often viewed as a genetic disease — one that results when the DNA sequence of key genes is mutated, leading to the removal of protective roadblocks and/or to the unchecked proliferation of cells. In recent years, however, it has become clear that 'epigenetic' malfunctions can also contribute to cancer development. These malfunctions can be defined as being due to relatively stable changes in gene expression without changes in the DNA sequence of the gene. They are often detected by the chemical addition or removal of methyl or acetyl groups that cause, or signal, a change in the structure of chromatin, the DNA-protein complex that forms chromosomes. When their normal modifying effects on gene expression are deregulated, tumours can arise.

Do some of the many 'surveillance' mechanisms that protect us from cancer act at the epigenetic level? One way to approach this problem is to ask whether mechanisms that influence the stringency of maintaining epigenetic modification can modulate the risk of cancer development. In a paper just out in *Science*¹, and in a previous study², Andrew Feinberg and colleagues address this issue.

The idea that changes in the epigenetic regulation of gene expression may contribute to the initiation or progression of tumour development was initially viewed with scepticism. But this view has changed radically, particularly in the past decade. Three kinds of frequent epigenetic changes have been identified in tumours. The first is genome-wide hypomethylation, which leads to the activation of many genes that are normally silenced in adult tissues. Second, tumour-suppressor genes can be inactivated by dense hypermethylation of their upstream regulatory sequences, with consequences that are similar to those resulting from inactivation of the same genes through mutation or deletion. Finally, loss of parental imprinting can occur. This is a specific form of epigenetic modification that involves inactivating either the maternally derived or the paternally derived copy of a particular gene in the offspring.

Given these findings, it would seem reasonable to suppose that organisms must have ways of keeping tabs on the epigenetic status of DNA, in the same way that other aspects of tumour development can be checked. At least three kinds of tumour surveillance have been firmly established³, and their importance is demonstrated by what happens

when they break down. The word 'surveillance' was first brought into the cancer field through the 'immune surveillance' hypothesis, which suggested that immune defence mechanisms nip precancerous cells in the bud. Later, however, it turned out that the immune system protects us mainly, if not exclusively, from the growth of potentially neoplastic cells that are initiated by viruses. The main examples include tumours associated with the human papillomaviruses, non-Hodgkin's lymphomas driven by Epstein-Barr virus, and Kaposi's sarcoma associated with human herpesvirus 8.

A second, phylogenetically more ancient and more powerful surveillance function is provided by the repair of damaged DNA sequences. The loss or impairment of DNA-repair enzymes is associated with multicancer syndromes. Finally, there is powerful intracellular surveillance, which detects the illegitimate activation of oncogenes (proliferation-driving genes) and can trigger programmed cell death. The crippling of major cell-death pathways is a universal feature of cancer.

So, is there also epigenetic surveillance? Feinberg and colleagues' findings^{1,2} — which indicate that variations in the stringency of parental imprinting may influence the likelihood of tumour development — suggest that there is. In their earlier paper², they found that loss of imprinting of the insulin-like growth factor II (*IGF2*) gene — a feature of many human cancers — can be detected in the peripheral blood lymphocyte cells of about 10% of the normal human population. This suggests that the stringency of the postnatal maintenance of *IGF2* imprinting is genetically determined; the authors also found that this 10% of the population has a 3.5–5-fold increase in the risk of developing colorectal adenomas; thus, instability of the mechanism for maintaining imprinting is related to the risk of neoplastic disease.

In their latest paper¹, the authors provide experimental evidence that supports their interpretation of the human data, and suggest a possible mechanism. In this study, hybrid mice were generated by crossing two genetically engineered mouse strains. The females used for the cross were heterozygous for a deleted differentially methylated region, or DMR (heterozygosity here means that one of the two copies of the relevant chromosome contained this region, whereas the other did not). Inheritance of this deletion from the mother leads to expression of

IGF2 from both maternally and paternally derived gene copies — thus corresponding to loss of imprinting. The males entering the hybrid cross came from the Min strain, which carries a mutation in the adenomatous polyposis coli (*APC*) gene. Mutations in *APC* provide a strong predisposition to familial colonic polyposis, a precancerous condition in humans and mice.

The result of this cross was that all offspring carried the *APC* mutation, but only half of them inherited the imprinting defect. Strikingly, the frequency of intestinal adenomas was twice as high in the mice with the imprinting defect compared with their littermate controls. Also, their intestinal crypts were longer and showed increased staining for proteins that are characteristic of intestinal-cell progenitors. This indicated that the differentiation of these cells to more specialized intestinal cell types was delayed.

These observations¹ show, for the first time, that the impairment of normal parental imprinting may interfere with cellular differentiation and thereby increase the probability of cancerous development. Taken together with the data in the earlier paper² — which showed person-to-person differences in the epigenetic silencing of *IGF2* that are related to a predisposition to colon cancer — the findings indicate that cancer susceptibility may be influenced by differences in the stringency of epigenetic control. This is consistent with earlier work showing that inbred mouse strains may differ in the activity of enzymes involved in DNA methylation⁴.

In more general terms, the findings^{1,2} also show that mutation of a 'cancer gene' (*APC*) and an epigenetically imposed delay in cell differentiation may drive tumour development synergistically. This inference is reminiscent of the early concept of two-phase carcinogenesis — by 'initiation' and 'promotion' — advanced by Isaac Berenblum⁵. In general, initiation may correspond to mutation or epigenetic silencing, whereas promotion probably occurs through epigenetic modification. More specifically, Berenblum showed that impairment of differentiation is an essential part of tumour promotion.

Feinberg and colleagues' results are consistent with this idea, but they also raise new questions. In particular, it will be necessary to explore the ways in which the loss of imprinting of a tumour-unrelated gene may contribute to the delay in differentiation of a specialized tissue, and influence its predisposition to cancer. ■

George Klein is in the Microbiology and Tumor Biology Center, Karolinska Institutet, 171 77 Stockholm, Sweden.
e-mail: Georg.Klein@mtc.ki.se

1. Sakatani, T. *et al. Science* doi:10.1126/science.1108080 (2005).
2. Cui, H. *et al. Science* **299**, 1753–1755 (2003).
3. Klein, G. *Cell Death Differ.* **11**, 13–17 (2004).
4. Paz, M. F. *et al. Cancer Res.* **62**, 4519–4524 (2002).
5. Berenblum, I. in *General Pathology* 2nd edn (ed. Florey, H.) (Saunders, Philadelphia, 1962).

Semiconductor technology

Negatively successful

Ananth Dodabalapur

Organic semiconducting polymers are promising electronic materials, but for full versatility they need to conduct negative as well as positive charge. A step towards that goal has now been taken.

Electronic devices based on organic materials are gradually becoming a mainstay of semiconductor technology, and are already used, for example, in flexible displays. However, a few hurdles remain before a breakthrough in practical applications can occur. One nagging problem is that organic transistors are so far only efficient at carrying electronic current via positive charge carriers (holes). To realize a complete electronic circuit based on organic polymers, analogous to conventional silicon chips, devices must also be able to carry negative charge (electrons). On page 194 of this issue, Chua *et al.*¹ demonstrate a design principle that would turn any organic field-effect transistor (FET) into one that efficiently carries negative charge.

The past decade has seen many efforts to develop so-called 'n-channel' organic FETs, in which electrons, rather than holes, are the charge carriers^{2–6}. This work is not aimed solely at immediate applications, but also at achieving a better understanding of the relation between materials and device performance.

The guiding principle in the design of semiconductor and electrode materials for n-channel FETs has been to optimize energy levels so as to achieve high electron mobility. But compared with 'p-channel' FETs (where the charge carriers are holes), such devices are generally found to be more sensitive to moisture and oxygen. To combat this problem, organic materials have been developed to have high electron affinity, which means that they strongly attract electrons and are not easily oxidized. However, electron affinity cannot be too high, or the material will become prone to unwanted doping, which is detrimental to the performance of a transistor.

An example of an organic semiconductor material based on this design principle of optimizing energy levels is a pair of materials known as F₁₅NTCDI and H₁₅NTCDI. Although the fluorine-free material has good electron mobility, transistors made with just this compound operate only under vacuum, where the effects of moisture and oxygen are much reduced. The fluorine-containing compound, on the other hand,

has a higher electron affinity and can operate in air².

Ultimately, materials are needed that are 'ambipolar' — that is, that can be used to create both n-channel and p-channel FETs. Until now, materials for making organic and polymer transistors have generally been lumped into one of the two categories, either n-channel or p-channel⁷. Traditional organic FET materials, such as pentacene, α -sexithiophene and polythiophenes, are known as p-channel materials and have been difficult to coerce into taking on n-channel behaviour.

However, experimental results suggest that the difficulty in observing efficient transport of electrons in an organic FET is an extrinsic effect caused by factors other than the organic semiconductor material itself. This conclusion would be in tune with the observation that, in purified single crystals of small organic semiconducting molecules, both electrons and holes move with roughly comparable mobilities.

For example, a traditional p-channel FET — based on pentacene — has been converted into an n-channel device by inserting calcium at the interface between pentacene and the thin insulating layer used in transistors⁸. The conclusion was that electrons from the calcium atoms fill up a large number of the 'traps' in the insulator that inhibit electron transport. Removal of these traps results in efficient n-channel behaviour.

Chua *et al.*¹ make a convincing case that the trapping of electrons at the insulator—

Structural biology

Methanol maker

Methanol could become an alternative source of energy, possibly replacing the less environmentally friendly combustion of coal and petroleum. The right-hand part of the picture shows burning methanol, which combusts much more cleanly than gasoline (on the left). Methanol can be produced by oxidizing methane, the main component of natural gas. But there are no catalysts that can efficiently perform this chemical reaction on an industrial scale at low temperature.

Several natural enzymes can carry out the reaction, however. Elsewhere in this issue (*Nature* **434**, 177–182; 2005), Raquel L. Lieberman and Amy C. Rosenzweig report the X-ray crystal structure of one of them — particulate methane monooxygenase (pMMO) from the bacterium *Methylococcus capsulatus*. The structure helps to settle years of controversy by

revealing that the protein is a cylindrical trimer, with three metal centres in each subunit. Two of the centres contain copper; one centre is mononuclear (containing a single copper ion), whereas the other is dinuclear (with two copper ions in close proximity). The other metal centre contains a zinc atom, which Lieberman and Rosenzweig believe came from the buffer used to crystallize the protein; they think that the natural protein contains either copper or iron at this centre.

It is still unclear which of the metal centres is the catalytic site — it could be any of them, say the authors. However, clues to the answer might come from structural similarities between a subunit of pMMO and cytochrome *c* oxidase subunit II, the terminal enzyme in the respiratory chain, which contains a dinuclear copper centre

involved in electron transfer. The dinuclear copper centre of pMMO could likewise be responsible for shuttling electrons to the catalytic site, which may be the nearby mononuclear copper centre.

The structure also shows which amino-acid side chains are associated with each metal ion. But much remains to be done. For example, the resolution of the structure is not high enough to determine whether the metal centres contain additional ligands, such as water, that could be functionally important. Nonetheless, the new work is a step along the way to finding out how pMMO works, and towards the long-term goal of developing



small-molecule catalysts that might replicate methane-to-methanol conversion in an industrial setting.

Joshua Finkelstein

J. BLAIR/CORBIS

semiconductor interface is indeed the culprit, and they relate this trapping to electronegative hydroxyl (OH) groups in the insulator material. When they use materials that are free of hydroxyl groups, uninhibited electron transport is indeed observed. By using bisbenzocyclobutene as the insulator, for example, organic polymer FETs based on materials such as poly(fluorine-*alt*-bithiophene) and poly(fluorine-*alt*-benzothiadiazole) are shown to exhibit n-channel behaviour.

Chua *et al.* conclude that if the trapping of electrons by electronegative groups in the insulating layer could be avoided, then n-channel behaviour would be seen in a broad range of semiconductors. Their most convincing example is that of polythiophene, which is relatively prone to oxidation and therefore normally more difficult to use in n-channel FETs.

Of course, it remains to be seen whether, on prolonged exposure of the device to ambient conditions, the effects of atmospheric moisture and oxygen would negate those of a hydroxyl-free insulator material. But the finding is nevertheless a remarkable result and a major step forward in our understanding of the design principles of materials and devices for n-channel organic transistors.

Chua and colleagues' work also explains why some semiconductors transport both electrons and holes when used as an active layer in light-emitting diodes — where there are no insulating interfaces — but exhibit only n-channel behaviour in FETs⁹. The implication is that such materials, which readily exhibit n-channel behaviour but not p-channel behaviour, could also be made ambipolar by identifying what causes hole trapping at the semiconductor–insulator interface and eliminating it.

Another area that will benefit greatly from this study is the work on ambipolar organic FETs with a view to making efficient light-emitting transistors¹⁰. A wider choice of materials for this type of research will become available once electron trapping is eliminated or greatly reduced.

Ananth Dodabalapur is in the Microelectronic Research Center, University of Texas at Austin, 10100 Burnet Road, Building 160, Austin, Texas 78758, USA.
e-mail: ananth@mer.utexas.edu

1. Chua, L.-L. *et al.* *Nature* **434**, 194–199 (2005).
2. Katz, H. E. *et al.* *Nature* **404**, 478–481 (2000).
3. Facchetti, A. *et al.* *J. Am. Chem. Soc.* **126**, 13859–13874 (2004).
4. Bao, Z. *Adv. Mater.* **12**, 227–230 (2000).
5. Chesterfield, R. J. *et al.* *J. Phys. Chem. B* **108**, 19281–19292 (2004).
6. Crone, B. *et al.* *Nature* **403**, 521–523 (2000).
7. Dimitrakopoulos, C. D. & Malenfant, P. R. L. *Adv. Mater.* **14**, 99–117 (2002).
8. Ahles, M., Schmechel, R. & Von Seggern, H. *Appl. Phys. Lett.* **85**, 4499–4501 (2004).
9. Ostrick, J. *et al.* *J. Appl. Phys.* **81**, 6804–6808 (1997).
10. Ahles, M., Hepp, A., Schmechel, R. & Von Seggern, H. *Appl. Phys. Lett.* **84**, 428–430 (2004).

Developmental biology

Sperm–egg fusion unscrambled

Richard Schultz and Carmen Williams

The identity of the sperm molecules that are involved in fusion with an egg's membrane has eluded biologists. Will Izumo, a protein named after the Japanese shrine to marriage, bring harmony to the field?

In the life history of sexually reproducing organisms, fertilization is a defining event. In mammals, sperm and egg have a limited lifespan once present in the female reproductive tract. Fertilization 'rescues' these gametes in a multi-step process that ultimately results in the fusion of their membranes and formation of a zygote. Despite the importance of the interactions that lead to zygote formation — and despite decades of research — the molecular basis for sperm–egg fusion has remained poorly understood, and the field is littered with the carnage of erstwhile molecular candidates. The latest discovery is of a sperm-specific protein called Izumo, and the evidence that it is required for fusion, presented by Okabe and colleagues¹ on page 234 of this issue, is compelling.

The steps involved in sperm–egg fusion are illustrated in Figure 1. Briefly, the cumulus cells that surround an ovulated egg secrete a matrix containing hyaluronic acid. A sperm must first pass through this matrix, and then interact with the egg's coat (the zona pellucida); this stimulates secretion of the contents of the sperm's acrosome — an intracellular sack of molecules necessary for sperm to penetrate the zona pellucida. Once through this coat, the sperm gains access to the perivitelline space, and binds to the egg's plasma membrane. Sperm and egg membranes then fuse.

Historically, antibodies have been used to identify proteins from the surface of mouse sperm that are involved in fertilization. In this way, PH-20 was identified as an enzyme that helps sperm to pass through the hyaluronic-acid-containing matrix². The same approach identified the fertilin $\alpha\beta$ protein as a candidate for mediating sperm–egg fusion — an exciting finding because of the structural nature of the fertilins. These proteins were the founding members of the ADAM (for 'a disintegrin and metalloprotease') superfamily³. Fertilin α and β each contain a disintegrin domain, which might mediate cell–cell adhesion by interacting with an egg integrin protein; fertilin α also contains a domain related to that of viral fusion peptides, which might mediate sperm–egg membrane fusion.

Unfortunately, however, later work did not support a role for fertilin α in membrane fusion, and sperm derived from fertilin- β -deficient mice can still fertilize eggs. Furthermore, wild-type sperm can bind to and fuse

with eggs deficient in each of the integrins that are likely to interact with ADAM disintegrin domains. These results deflated the previous excitement and left the field in a profound state of uncertainty.

The results described in this issue by Okabe and colleagues¹ should rekindle a sense of vitality in this area of research. Okabe and colleagues⁴ reported 18 years ago that a particular antibody, OBF13, recognizes a molecule on the sperm head that is exposed following the acrosome reaction. A later study⁵ indicated that although this antibody did not inhibit sperm–egg binding, it did prevent fertilization *in vitro*, hinting that the protein it binds is involved in fusion. Now, Okabe and colleagues¹ have used liquid chromatography tandem mass spectrometry to identify this protein, and have named it Izumo. The protein sequence indicates that Izumo (and its human counterpart) is a member of the immunoglobulin superfamily, and other experiments show it to be sperm-specific.

The authors used a knockout approach to assess a role for Izumo in fertilization. Females lacking the protein were phenotypically normal and fertile. In contrast, although Izumo-deficient males were phenotypically normal — and their sperm were also normal in morphology and motility — they were completely infertile. The Izumo-deficient sperm could reach eggs *in vivo*. They showed no reduction in their ability to bind to and penetrate the zona pellucida *in vitro*. And, once in the perivitelline space, they could bind to the egg plasma membrane. But not a single Izumo-deficient sperm was observed to fuse with the egg membrane. The authors ascribe this inability directly to the lack of Izumo, because expressing this protein in Izumo-deficient males restored fertility. Finally, they found that bypassing membrane fusion by injecting Izumo-deficient sperm into eggs resulted in activation of the eggs and development to term. These results, and more, strongly suggest that the sole function of Izumo is in sperm–egg fusion.

So, how does Izumo promote membrane fusion? Because its extracellular region lacks sequences like those found in viral fusion peptides, it is unlikely that Izumo is inherently fusogenic. A more likely possibility is that it serves as an adhesion molecule, given that it has an immunoglobulin-like domain — a well-defined domain for mediating cell–cell adhesion (see, for example, ref. 6). In contrast to most members of the

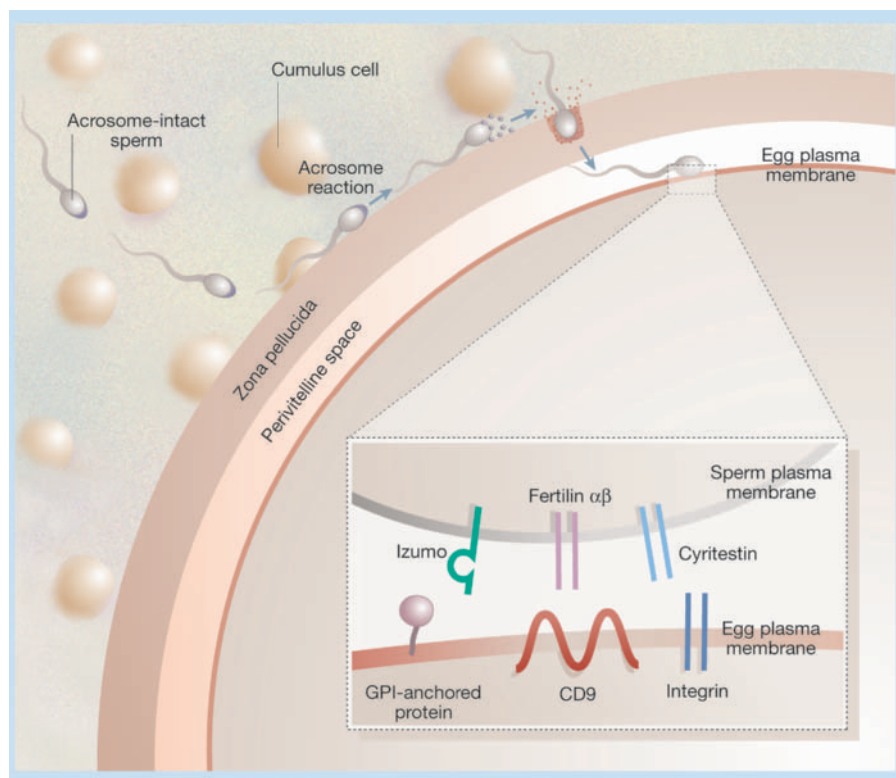


Figure 1 Events leading to fertilization. A sperm (whose acrosome compartment is intact at this point) passes through the hyaluronic-acid-containing matrix secreted by cumulus cells. Interaction with the egg's coat (the zona pellucida) leads to the secretion of acrosomal contents and exposure of molecules needed for sperm-egg binding and fusion. Although the zona pellucida is well characterized biochemically (it is composed of several glycoproteins), the molecular basis for how sperm bind to it and undergo the acrosome reaction is not resolved¹². The sperm then penetrates the zona pellucida, gains access to the perivitelline space and binds to the egg's plasma membrane. Inset, molecules on the sperm surface, such as fertilin and cyritestin³, may be involved in sperm-egg binding; Okabe and colleagues¹ find that Izumo is essential for membrane fusion. On the egg, CD9 is required for fusion and might collaborate with other proteins such as integrins or glycosylphosphatidylinositol (GPI)-anchored proteins.

immunoglobulin superfamily, the relatively small extracellular region of Izumo contains only a single immunoglobulin-like domain. But, by interacting with an egg surface protein, this small domain could permit the close membrane apposition requisite for membrane fusion.

Assuming that Izumo binds to egg proteins, what might they be? So far, the tetraspanin-superfamily protein CD9 is the only egg protein documented to function in sperm-egg fusion^{7,8}. Tetraspanins mediate protein interactions to create unique cell-surface regions that contain specific proteins, including integrins, immunoglobulin-superfamily members and other tetraspanins. Female mice without CD9 are phenotypically normal but infertile. Sperm bind normally to CD9-deficient eggs but fail to fuse; this defect is eliminated by injecting CD9-encoding messenger RNA⁹. Because CD9 can interact with other immunoglobulin-superfamily members — such as pregnancy-specific glycoprotein 17 (ref. 10) — it is possible that it also interacts directly with Izumo. Alternatively, a CD9-associated protein, perhaps a glycosylphosphatidylinositol-anchored protein, might participate; sperm do not fuse with eggs lacking such proteins¹¹. Crosslinking experiments, using Izumo as a probe, would be an obvious way to try to identify potential partners in the egg, as well as sperm proteins that interact with Izumo.

Finally, although most of Okabe and colleagues' work¹ was conducted using mice, they also note that humans express a counterpart of Izumo on the sperm head, and that an antibody raised against this

protein blocks the fusion of human sperm with hamster eggs — a test for male fertility. These findings raise the exciting possibility that, because Izumo is sperm-specific and extracellular, this protein and its interacting partners could be new targets for non-hormonal contraception.

Richard Schultz is in the Department of Biology and Carmen Williams is at the Center for Research on Reproduction and Women's Health, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
e-mail: rschultz@mail.sas.upenn.edu

Planetary sciences

A smashing pair

Jay Melosh

The likely origin of Pluto and its satellite Charon, like the Earth and Moon, is an impact between two planet-sized bodies. Refined simulations show that there may be two distinct modes for the birth of such twins.

Since Galileo's unexpected discovery of Jupiter's four large moons in 1610, astronomers have continued to discover satellites in strange places, with the latest found among the distant, icy objects in the Kuiper belt. However, our understanding of how these satellites came into existence has not kept pace with discovery. Indeed, the origins of the satellites of Mars and the irregular satellites of Jupiter are still fraught with uncertainty.

The origin of our own Moon was similarly shrouded in confusion until the

mid-1980s, when a new and dramatic idea, the giant-impact hypothesis, swept all others before it¹. During the heated debates about the origin of Earth's moon, it was suggested that the Solar System's other planetary twins, Pluto and Charon, might also be the offspring of a gigantic collision². Now Robin Canup, the leading researcher into simulations of the terrestrial giant impact, has applied her computer models to the Pluto system and added detail and precision to the violent birth of Pluto's moon³.

Unlike the rather loosely bound systems

1. Inoue, N., Ikawa, M., Isotani, A. & Okabe, M. *Nature* **434**, 234–238 (2005).
2. Hunnicutt, G. R., Primakoff, P. & Myles, D. G. *Biol. Reprod.* **55**, 80–86 (1996).
3. Stein, K. K. et al. *J. Cell Sci.* **117**, 6269–6274 (2004).
4. Okabe, M. et al. *J. Reprod. Immunol.* **11**, 91–100 (1987).
5. Okabe, M. et al. *J. Reprod. Immunol.* **13**, 211–219 (1988).
6. Taylor, M. V. *Curr. Biol.* **12**, R224–R228 (2002).
7. Le Naour, F., Rubinstein, E., Jasmin, C., Prenant, M. & Boucheix, C. *Science* **287**, 319–321 (2000).
8. Miyado, K. et al. *Science* **287**, 321–324 (2000).
9. Kaji, K. et al. *Dev. Biol.* **247**, 327–334 (2002).
10. Waterhouse, R., Ha, C. & Dveksler, G. S. *J. Exp. Med.* **195**, 277–282 (2002).
11. Alfieri, J. A. et al. *J. Cell Sci.* **116**, 2149–2155 (2003).
12. Dean, J. *BioEssays* **26**, 29–38 (2004).

of satellites in the Kuiper belt or among the asteroids, the Pluto–Charon system is a tightly bound pair whose angular momentum, if concentrated in a single object with the combined mass of the two, would produce an object rotating so fast that it approaches rotational break-up. This is similar to the Earth–Moon system and is an immediate consequence of an origin by impact: the parent object must have been a fleeting association between two planet-sized bodies that lost a substantial fraction of their initial energy during a brief, violent collision. The main result of such an impact is that a portion of the gravitational and kinetic energy of the bodies is converted into heat, robbing them of the ability to go their separate ways. As Canup found in her detailed simulations, this process is most effective when the relative velocity of the two approaching bodies is small, thus minimizing the difference in total energy between a bound pair and two separate bodies.

The major constraints on the collision process are the final angular momentum of the planet–satellite pair and the ratio of their masses. Whereas our Moon's mass is about 1% of Earth's, that of Charon is a huge 10–15% of Pluto's mass. Unfortunately, there are too many unknown factors to determine the precise conditions of the impact: the mass, structure and initial rotation states of both the participants in the collision, their relative velocity and the angle of impact. Canup's procedure is thus probabilistic by necessity. She examined a broad suite of different possible initial conditions and evaluated how many of these produced a final configuration similar to the Pluto–Charon system.

Her simulations use a method called smooth particle hydrodynamics, which incorporates a full three-dimensional geometry. In addition, the gravitational forces between each portion of both distorted planets must be taken into account. In agreement with earlier researchers⁴, Canup found that torques exerted by the distorted mass that is the product of the collision are crucial for inserting a large amount of the mass into orbit to produce a moon. What is new about this work is that she discovered two basic scenarios for the creation of a satellite.

In the first scenario, the parent bodies are differentiated, which means they have dense cores overlain by a lighter mantle. Most of their mass merges into a new central body, while gravitational torques spin the remaining material out into a dispersed ring or disk. The satellite later condenses from the disk. This disk is mainly composed of intensely heated mantle material, and the ratio of the disk mass to the planet's mass is relatively small. This may be the process that created our Moon.

Undifferentiated bodies, on the other hand, which do not have a distinct core and mantle, remain more intact after collision.

Animal behaviour

Meals sized up

Termites are fussy eaters. A particular piece of wood may be avoided because it is too hard or contains defensive chemicals that the termites can't detoxify. Different termite species also vary in their food requirements, and such selectivity may help to reduce competition, allowing different species to coexist in the same habitat. Even wood size matters — one species may specialize on entire timber-framed buildings, whereas another may have an appetite only for twigs.

Theodore A. Evans and colleagues have tackled the problem of how termites assess wood size (*Proc. Natl Acad. Sci. USA* **102**, 3732–3737; 2005). Worker drywood termites (*Cryptotermes domesticus*, pictured) are blind and so



cannot judge wood size visually and do not survey the wood physically before starting to eat. The researchers considered another possibility. Termites often communicate with vibrational signals — soldiers, for example, may drum their heads against the wood to warn of impending danger — and they are also noisy eaters.

Evans *et al.* wondered whether termites use vibrations generated during foraging to judge the resonance frequency of the wood, which is related to

its size. To test this possibility, they presented hungry worker termites with blocks of wood of different sizes. The workers preferred blocks of a particular size, but this preference could be specifically altered by playing them recordings of termites feeding, or by producing artificially generated vibrations. Remarkably, the vibrational signals also affected reproductive development in the species, suggesting that such signals might play a wider role in termite biology than has been appreciated.

Rory Howlett

Although they again share angular momentum through gravitational torques, the satellite is essentially complete at birth. The ratio of satellite mass to final planet mass is much higher, and the bulk of the satellite is not strongly heated. This is the likely circumstance of the birth of Pluto and Charon. Each object retains nearly its initial chemical composition. Such a scenario would be a poor candidate for the Earth–Moon system because the Moon differs substantially from the Earth, especially in its content of metallic iron–nickel, and therefore its density. We know little of the composition of Pluto and Charon (current measurements⁵ are not able to clearly discriminate between planet and moon), but the prediction that they are very similar in density can be tested by future spacecraft reconnaissance or by improvements in astrometric measurements.

What is the origin of the object that became Charon? Canup's simulations do not answer this question, but they do offer some clues. The need for a low approach velocity suggests that the two parents were in very similar initial orbits. Further, the present Pluto–Charon system is in a rather delicate resonance with Neptune that prevents a collision between Neptune and Pluto–Charon in spite of their crossing orbits: Neptune circles the Sun exactly 1.5 times faster than Pluto. Any scenario describing the giant impact that produced Pluto and Charon must yield a system that ends up in this resonance.

After evaluating a number of different possibilities, Canup concludes that the two

parents could already have been trapped in a resonance with Neptune before colliding with each other. Trapping may have occurred during a time when the outer planets' orbits were evolving through interactions with a massive disk that became the Kuiper belt. As Neptune's orbit expanded, its orbital period would also have increased. The orbital periods of the parent bodies gradually synchronized with Neptune, trapping them in the stable 1.5 resonant configuration⁶. They remained in this resonance during and after the collision that produced the Pluto–Charon system.

As with all numerical simulations, future work at higher resolution and with better thermodynamic descriptions of the materials making up the colliding objects is always desirable. However, Canup has added a substantial amount of flesh to the skeleton of the giant-impact idea and made this catastrophic interpretation of the Pluto–Charon system far more plausible. Her elucidation of two distinct modes of satellite formation in large impacts has added reason and clarity to our ever-improving understanding of these processes.

Jay Melosh is in the Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA. e-mail: jmelosh@lpl.arizona.edu

1. Canup, R. M. *Annu. Rev. Astron. Astrophys.* **42**, 441–475 (2004).
2. McKinnon, W. B. *Nature* **311**, 355–358 (1984).
3. Canup, R. M. *Science* **307**, 546–550 (2005).
4. Cameron, A. G. W. & Benz, W. *Icarus* **92**, 204–216 (1991).
5. Olkin, C. B., Wasserman, L. H. & Franz, O. G. *Icarus* **164**, 254–259 (2003).
6. Malhotra, R. *Nature* **365**, 819–821 (1993).

Palaeobiology

Dating earliest life

Stephen Moorbath

Claims that 3.8-billion-year-old rocks from Greenland contain carbonaceous remnants of very early life have been the subject of argument for several years. The latest analyses look like settling matters.

How, where and when did life start? Geologists are helping to address these questions by diligently searching the oldest-known sedimentary rocks on Earth for traces of primitive life, either in the form of cellular microfossils, or as chemical and isotopic tracers characteristic of biological processes.

Since 1996, attention has focused on a small outcrop of supposedly sedimentary rocks, claimed to be more than 3.85 billion years old, on the tiny island of Akilia, just off the west coast of Greenland. These rocks have been too strongly recrystallized in subsequent geological eras by heat and pressure (metamorphism), penetration of deep crustal fluids (metasomatism) and tectonic deformation to preserve fossils. Hence, claims for life there had to be based on the carbon isotope composition of tiny inclusions of graphite (carbon) in grains of apatite (calcium phosphate). Writing in *Geology*, Lepland *et al.*¹ now report that they cannot substantiate claims for the presence of graphite in any of the crucial Akilia apatite grains.

The story began with the publication of two papers^{2,3} claiming that ¹³C isotope depletion (that is, low ¹³C/¹²C ratios) in Akilia graphite supported a biological origin. The host rocks were identified as a banded iron formation (BIF), a chemically precipitated marine sediment normally consisting of alternating bands of iron oxide (magnetite or haematite) and silica (quartz) that might, in principle, harbour remnants of early life. Such rocks are difficult to date directly, but an age of 3.85 billion years or more was claimed from uranium–lead dating of zircon (zirconium silicate) grains in a granitic vein cross-cutting — and therefore younger than — the rock that hosted the BIF. This date provided ammunition for the widespread popular claim that life on Earth began around 4 billion years ago. It also implied that earliest life coexisted with devastating global meteorite impacts, known as the Late Heavy Bombardment, which shaped the Moon's surface and, most likely, that of the Earth until about 3.85 billion years ago.

Following the appearance of the two papers^{2,3}, scientific debate commenced in earnest. I myself visited Akilia twice, in company with several participants in the unfolding controversy, but I could not understand why these banded rocks had been identified

as BIF. Instead of iron oxide, the dark bands consisted of pyroxene (an aluminosilicate of calcium, magnesium and iron) and resembled igneous rocks from very close by. Genuine BIF occurs in quantity some 150 km to the northeast in the famous Isua region, where the best-preserved 3.7–3.8-billion-year-old rocks on Earth occur. The Akilia rocks looked nothing like the Isua BIFs. Detailed chemical and mineralogical evidence soon convinced Fedo and Whitehouse⁴ that the rocks were not sediments but strongly metamorphosed, metasomatized and deformed igneous rocks, and thus irrelevant for biology.

The zircon uranium–lead date of 3.85 billion years or older for the Akilia rocks was also questioned, because individual zircon grains are complex, possessing zones of quite different ages within a single grain that represent its long and complex crystallization history. The oldest zones indeed gave nearly 3.85 billion years, but this could be an 'inherited' age, older than the age of the host granitic vein, which some claimed to be no more than 3.65 billion years⁵. Other workers⁶ then objected that the granitic vein was not cross-cutting at all, but structurally concordant with its adjacent rock. That meant that the measured age of the granitic vein was irrelevant to the crucial Akilia rock, because any previous field relationship between the two had been obliterated by tectonic forces.

Even more worrying was the proposal⁷ that graphite inclusions in apatite grains resulted from thermal dissociation (at temperatures of about 450 °C) of iron-rich carbonates in the rock to iron oxide, carbon dioxide and elemental carbon, the latter with a ¹³C/¹²C ratio overlapping values indicative of a biogenic origin. This mechanism was suggested to counter claims for supposedly biogenic carbon in the genuine BIFs from Isua⁸, which often contain iron-rich carbonate. In fact, only one locality now remains in the Isua region where association of graphite with sedimentary rocks leaves the possibility of a biogenic interpretation open⁹. Exchanges between 'pro-life' and 'anti-life' factions on Akilia have now been published for eight years. However, for many observers the ancient Akilia rocks have gradually appeared as less and less plausible domiciles for biogenic tracers.

Lepland *et al.*¹ investigated many thin sections of Akilia rocks containing apatite

crystals by optical and scanning electron microscopy, combined with energy-dispersive spectrometry. This work failed to reveal any graphite inclusions in apatite crystals from the problematic banded rocks or surrounding rocks. Even the most widely discussed sample (known as G91-26), used in the original study², proved free of graphite.

This persuasive discovery seems an almost inevitable, yet highly problematic, consequence to the increasing scientific doubts about the original claim. We may well ask what exactly was the material originally analysed and reported? What was the apatite grain with supposed graphite inclusions that figured on the covers of learned and popular journals soon after the discovery? These questions must surely be answered and, if necessary, lessons learned for the more effective checking and duplication of spectacular scientific claims from the outset.

To my regret, the ancient Greenland rocks have not yet produced any compelling evidence for the existence of life by 3.8 billion years ago. The reader is reminded that another debate on early life is currently in progress on 3.5-billion-year-old rocks in Western Australia, where chains of cell-like structures, long identified as genuine fossils¹⁰, have recently been downgraded by some workers¹¹ to the status of artefacts produced by entirely non-biological processes. To have a chance of success, it seems that the search for remnants of earliest life must be carried out on sedimentary rocks that are as old, unmetamorphosed, unmetasomatized and undeformed as possible. That remains easier said than done. For the time being, the many claims for life in the first 2.0–2.5 billion years of Earth's history are once again being vigorously debated: true consensus for life's existence seems to be reached only with the bacterial fossils of the 1.9-billion-year-old Gunflint Formation of Ontario¹².

Stephen Moorbath is in the Department of Earth Sciences, University of Oxford, Parks Road, Oxford OX1 3PR, UK.

e-mail: stephenm@earth.ox.ac.uk

1. Lepland, A., van Zuilen, M. A., Arrhenius, G., Whitehouse, M. J. & Fedo, C. M. *Geology* **33**, 77–79 (2005).
2. Mojzsis, S. J. *et al.* *Nature* **384**, 55–59 (1996).
3. Nutman, A. P., Mojzsis, S. J. & Friend, C. R. L. *Geochim. Cosmochim. Acta* **61**, 2475–2484 (1997).
4. Fedo, C. M. & Whitehouse, M. J. *Science* **296**, 1448–1452 (2002).
5. Whitehouse, M. J., Kamber, B. S. & Moorbath, S. *Chem. Geol.* **160**, 201–224 (1999).
6. Whitehouse, M. J. & Fedo, C. M. *Precamb. Res.* **126**, 259–271 (2003).
7. van Zuilen, M. A., Lepland, A. & Arrhenius, G. *Nature* **418**, 627–630 (2002).
8. van Zuilen, M. A. *et al.* *Precamb. Res.* **126**, 331–348 (2003).
9. Rosing, M. T. & Frei, R. *Earth Planet. Sci. Lett.* **217**, 237–244 (2004).
10. Schopf, J. W. *Science* **260**, 640–646 (1993).
11. Brasier, M. D. *et al.* *Nature* **416**, 76–81 (2002).
12. Knoll, A. H. *Life on a Young Planet* (Princeton Univ. Press, 2003).

Neurobiology

Agents against cocaine addiction

J. Neurosci. **25**, 1889–1893 (2005)

The search for treatments for cocaine addiction has been frustrating. Most notably, many of the compounds that bind to the drug's apparent biological targets, dopamine transporters, induce cocaine-like behavioural effects.

However, analogues of a drug called benztropine are showing promise as potential medications. Rajeev I. Desai and colleagues now show that one such analogue, JHW007, has a high affinity for dopamine transporters but not the same stimulatory effects as cocaine. When cocaine itself binds to these transporters, it prevents the reuptake of dopamine; this neurotransmitter therefore accumulates and begins to activate other cells, creating an intense 'high'. The researchers think that the ten times slower binding of JHW007 to dopamine transporters is responsible for its lack of cocaine-like effects.

Moreover, JHW007 has the virtue of crossing the blood–brain barrier, which means that it reaches the brain comparatively readily and, once there, it antagonizes the stimulant effects of cocaine. As well as pointing to the attributes of this analogue as a potential treatment for cocaine addiction, Desai *et al.* mention the scale of the cocaine problem: according to some surveys, there are more than two million cocaine users in the United States alone.

Roxanne Khamsi

Developmental biology

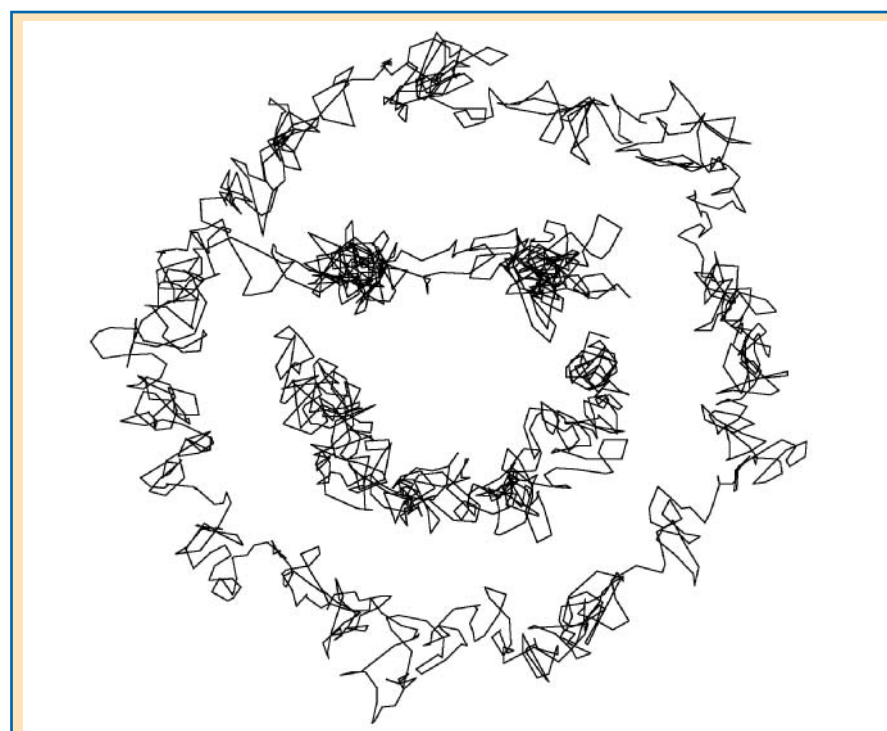
DiGeorge mice

Genes Dev. **19**, 530–535 (2005)

People with a rare congenital condition called DiGeorge syndrome have an array of facial, heart, endocrine and immune abnormalities. Heiko Wurdak *et al.* show how problems with signalling by a molecule called transforming growth factor- β (TGF- β) may contribute to this condition.

The researchers examined the role of TGF- β in neural-crest cells, which migrate from the neural tube in the growing embryo to various locations that are affected by DiGeorge syndrome, including the face, heart and thymus. They found that mice genetically engineered to lack TGF- β in their neural-crest cells mimic many of the defects seen in DiGeorge syndrome patients. Their observations suggest that the cells are unable to adopt the fates of smooth muscle or other tissues.

Wurdak *et al.* propose that TGF- β partly controls the activity of another protein, called CrkL, which lies in a segment of chromosome 22 that is missing in



Nanotechnology

Stop for brownian motion

Appl. Phys. Lett. **86**, 093109 (2005)

Making nanoparticles is one thing; controlling them in solution is another. Laser tweezers, alternating electric currents and magnetic fields can all be used to trap and manipulate micrometre-sized objects, but fail for much smaller particles.

Adam E. Cohen and W. E. Moerner have an answer that can raise a smile. They have used a direct electrical current hooked up to a digital feedback system to cancel out the brownian motion of a nanoparticle in aqueous solution, allowing them to trap fluorescent polystyrene nanospheres as small as 20 nm across.

As the particle moves around, a fluorescence microscope monitors its progress and alters the voltage across the solution to correct that drift. This keeps the particle locked into place, while careful variation of the voltage allows the particle to be moved at will — hence the smiley face shown here, which was drawn with a 200-nm-diameter particle and is about 10 μm across.

The trap should work for any object that can be imaged optically and acquires a charge in water, the authors say. With the present set-up the particle can be confined to within 1 μm . But an improved apparatus can offer increased precision and power, as they will report in a forthcoming paper in *Physical Review Letters*.

Mark Peplow

DiGeorge patients. Thus, they speculate that at least part of the syndrome is caused when loss of this gene disrupts TGF- β signalling.

Helen Pearson

Metrology

Requiem for a mass?

Metrologia **42**, 71–80 (2005)

One would think that objects kept in a vault would not gather much dust. Not so. The platinum–iridium prototype of the standard kilogram must be thoroughly cleaned before each use and, despite this, its mass may have varied by as much as 100 μg (about five grains of sugar) over the past century.

Redefining the kilogram using a fundamental constant such as Planck's constant h or Avogadro's constant N_A could eliminate this source of uncertainty. Ian M. Mills and co-workers propose to redefine the kilogram by fixing the value

of either h or N_A , rather like fixing the speed of light to define the metre.

Currently, h and N_A are themselves known to within only about 10^{-7} (one part in ten million), which is similar to the uncertainty in the mass of the prototype kilogram. Previous proposals for a new standard kilogram have always suggested waiting until the accuracy of h and N_A has been improved by a factor of at least ten.

But the authors observe that three advantages will follow from fixing h or N_A now: the kilogram will be defined in terms of an 'invariant of nature', rather than an unstable artefact; the uncertainties of many other fundamental constants (such as the elementary charge e and electron and proton masses m_e and m_p) would be reduced by more than a factor of ten; and future changes in the best estimates of the fundamental constants from one review to the next would be markedly reduced.

May Chiao

Meteor Crater formed by low-velocity impact

The paucity of melted rock in this crater may be due to the striking projectile's speed.

Meteor Crater in Arizona was the first terrestrial structure to be widely recognized as a meteorite impact scar and has probably been more intensively studied than any other impact crater on Earth. We have discovered something surprising about its mode of formation — namely that the surface-impact velocity of the iron meteorite that created Meteor Crater was only about 12 km s^{-1} . This is close to the 9.4 km s^{-1} minimum originally proposed¹ but far short of the $15\text{--}20 \text{ km s}^{-1}$ that has been widely assumed² — a realization that clears up a long-standing puzzle about why the crater does not contain large volumes of rock melted by the impact.

The Earth's atmosphere is an effective but selective screen against small meteoroids: it admits strong, dense iron projectiles before it permits stony objects to penetrate. Large projectiles of both types do penetrate the atmosphere, although they may be fractured by aerodynamic stresses on the way to the surface. Detailed, sophisticated analysis of this process has produced a variety of successful models for the cascade of fragmentation and dispersion of entering projectiles^{3–6}.

We applied one of the simpler models⁴ (see supplementary information for methods)

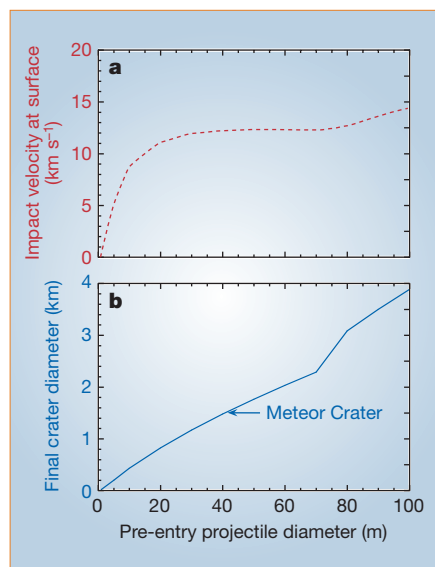


Figure 1 Surface-impact velocity and final crater size as a function of pre-atmospheric projectile size. **a**, Impact velocity. **b**, Crater diameter. The pre-atmospheric impact velocity is assumed to be 17 km s^{-1} , the angle of approach is 45° and the iron projectile's density is $8,000 \text{ kg m}^{-3}$. Its crushing strength is 50 MPa , in agreement with estimates based on observed meteorite falls, such as that of Shikote-Alir⁵. The computations assume a pancake model⁴ with a debris cloud expansion factor of 4, followed by separation of a fragment that is half the mass of the original projectile, which traverses the final few kilometres as an intact fragment.



Making an impression: Meteor Crater in Arizona ($35^\circ 2' \text{ N}$, $111^\circ 1' \text{ W}$) is roughly 1.2 kilometres across.

to Meteor Crater. Our results revealed that the projectile would have been greatly slowed by the atmosphere and would have struck as a dispersed cluster of iron fragments. We conclude that the fragmented iron projectile probably struck the surface at a velocity of about 12 km s^{-1} . Figure 1 shows the surface-impact velocity and final crater size for a suite of spherical iron projectiles of up to 100 m in diameter.

In our scenario, atmospheric drag slows the incoming projectile. As the density of the atmosphere increases, the stagnation pressure reaches 'crushing strength' at an altitude of about 14 km . After the meteor breaks up, the fragments spread out and the drag force on the cluster increases dramatically, resulting in a strong positive feedback. By an altitude of about 5 km , the cluster has expanded to a cloud (shaped like a pancake) of about 200 m in diameter, with a velocity of 13 km s^{-1} . At this point, we assume that a fragment that is one-half the mass of the original projectile separates and continues intact to the surface, where it strikes at 12 km s^{-1} , releasing about 2.5 MT (megatonnes of TNT equivalent) of energy. This is a conservative assumption because the aerodynamic stresses are much larger than the crushing strength of iron, and further episodes of crushing and deceleration should occur. The remainder of the projectile's initial energy, about 6.5 MT , is deposited in the atmosphere and initiates a strong airblast.

The surface-impact velocity is too low for substantial melting of the target rock⁷. This result may explain the old observation that there appears to be much less melt in Meteor Crater than would be expected by

extrapolation from larger craters⁸. The standard explanation for this discrepancy has been that the porous sedimentary target rocks contained groundwater and this water dispersed the melt into tiny droplets as it vaporized, or that carbonates in the target decomposed explosively to yield carbon dioxide⁸. However, if the consequences of atmospheric entry are properly taken into account, it appears that there is no melt discrepancy at all. Our proposal that Meteor Crater was created by a dispersed swarm of low-velocity iron fragments (many of which were dispersed beyond the central debris cloud) is also consistent with the recovery of large numbers of small, unmelted iron-meteorite fragments near the crater⁹.

H. J. Melosh, G. S. Collins*

Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA

e-mail: jmelosh@lpl.arizona.edu

*Present address: Earth Science and Engineering, Imperial College, London SW7 2AZ, UK

1. Shoemaker, E. M. in *The Moon, Meteorites and Comets* (eds Middlehurst, B. M. & Kuiper, G. P.) 301–336 (Univ. Chicago Press, Chicago, 1963).
2. Bryan, J. B., Burton, D. E., Cunningham, M. E. & Lettis, L. A. in *Proceedings of the Ninth Lunar and Planetary Science Conference* 3931–3964 (Pergamon, Houston, Texas, 1978).
3. Passey, Q. & Melosh, H. J. *Icarus* **42**, 211–233 (1980).
4. Chyba, C. F., Thomas, P. J. & Zahnle, K. J. *Nature* **361**, 40–44 (1993).
5. Svetsov, V. V., Nemtchinov, I. V. & Teterev, A. V. *Icarus* **116**, 131–153 (1995).
6. Bland, P. A. & Artemieva, N. A. *Nature* **424**, 288–291 (2003).
7. Pierazzo, E., Vickery, A. M. & Melosh, H. J. *Icarus* **127**, 408–432 (1997).
8. Kieffer, S. W. & Simonds, C. H. *Rev. Geophys. Space Phys.* **18**, 143–181 (1980).
9. Ninninger, H. H. *Arizona's Meteorite Crater* (American Meteorite Laboratory, Denver, Colorado, 1956).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Musical imagery

Sound of silence activates auditory cortex

Auditory imagery occurs when one mentally rehearses telephone numbers or has a song ‘on the brain’ — it is the subjective experience of hearing in the absence of auditory stimulation, and is useful for investigating aspects of human cognition¹. Here we use functional magnetic resonance imaging to identify and characterize the neural substrates that support unprompted auditory imagery and find that auditory and visual imagery seem to obey similar basic neural principles.

The few studies that have examined the topic of auditory imagery^{2–5} have focused on the neural substrates of directed imagery (for example, “imagine a tone”). What is not known, however, is whether similar principles guide the more pervasive and spontaneous forms of imagery that punctuate everyday life. We used functional magnetic resonance imaging to investigate the recruitment of auditory cortex during spontaneous auditory imagery of excerpts of popular music.

During scanning, subjects passively listened to excerpts of songs with lyrics (for example, *Satisfaction* by the Rolling Stones) and to instrumentals that contained no lyrics (for example, the theme from *The Pink Panther*). Each piece of music was pre-rated by subjects as either familiar or unknown, and a unique soundtrack was created for each individual. Short sections of music (lasting for

2–5 s) were extracted at different points during the soundtrack and replaced with silent gaps. We then monitored the neural activity in subjects that occurred during these gaps. (For details of methods, see supplementary information.)

Brain activity in the primary auditory cortex and in the auditory association cortex (Brodmann’s area 22) (Fig. 1a) was compared during gaps of silence in familiar and unknown songs. The results revealed a functional dissociation within the left auditory cortex (region $\times$ music-type interaction: $F[1,14] = 48.92$, $P < 0.0001$; Fig. 1b). Silent gaps embedded in familiar songs induced greater activation in auditory association areas than did silent gaps embedded in unknown songs (Fig. 1b); this was true for gaps in songs with lyrics ($F[1,14] = 5.46$, $P < 0.05$; Fig. 1c) and without lyrics ($F[1,14] = 11.56$, $P < 0.005$; Fig. 1d). Moreover, when familiar songs contained no lyrics, cortical activity extended into the left primary auditory cortex ($F[1,14] = 22.55$, $P < 0.0005$; Fig. 1d).

We confirmed that these effects were uniquely attributable to the gaps of silence in the music, rather than simply the result of differences in activation in response to hearing different music categories. By contrast with the gap responses, listening to unknown songs produced greater activity in auditory association areas than did familiar songs (lyrics: $F[1,14] = 11.24$, $P < 0.005$; instrumentals: $F[1,14] = 31.74$, $P < 0.0001$), and activity in the primary auditory cortex did not differ as a function of familiarity (see supplementary information).

Our findings offer a neural basis for the spontaneous and sometimes vexing experience of hearing a familiar melody in one’s head. Whereas previous investigations have explicitly directed subjects to imagine a specific auditory experience^{2–4}, we provided no instruction. Instead, simply muting short gaps of familiar music was sufficient to trigger auditory imagery — a finding that indicates the obligatory nature of this phenomenon. Corroborating this observation, all subjects reported subjectively hearing a continuation of the familiar songs, but not of the unfamiliar songs, during the gaps in the music.

We note also that the extent of neural activity in the primary auditory cortex was determined by the linguistic features of the imagined experience. When semantic knowledge (that is, lyrics) could be used to generate the missing information, reconstruction terminated in auditory association areas. When this meaning-based route to reconstruction was unavailable (as in instrumentals), activity extended to lower-level regions of the auditory cortex, most notably the primary auditory cortex (Fig. 1b,d).

These findings parallel those in the domain of visual imagery. For example, visual imagery elicited when considering names of objects (known as figural imagery) does not rely on the primary visual cortex^{6,7}. As these ‘low-resolution’ images do not demand fine-grained perceptual processing, activity in visual-association areas is sufficient to reconstruct the relevant representation. By contrast, when semantic information is absent or irrelevant (known as depictive imagery), a ‘high-resolution’ perceptual image is needed to reconstruct a representation, hence activity extends into the primary visual cortex⁸. Our results provide evidence that auditory imagery obeys the same basic neural principles.

David J. M. Kraemer*, C. Neil Macrae*†, Adam E. Green*, William M. Kelley*

*Department of Psychological and Brain Sciences, Dartmouth College, Hanover, New Hampshire 03755, USA

e-mail: william.kelley@dartmouth.edu

†School of Psychology, University of Aberdeen, Aberdeen AB24 2UB, UK

1. Kosslyn, S. M., Ganis, G. & Thompson, W. L. *Nature Rev. Neurosci.* **2**, 635–642 (2001).
2. Halpern, A. R. & Zatorre, R. J. *Cereb. Cortex* **9**, 697–704 (1999).
3. Wheeler, M. E., Petersen, S. E. & Buckner, R. L. *Proc. Natl Acad. Sci. USA* **97**, 11125–11129 (2000).
4. Yoo, S. S., Lee, C. U. & Choi, B. G. *Neuroreport* **12**, 3045–3049 (2001).
5. Hughes, H. C. *et al. Neuroimage* **13**, 1073–1089 (2001).
6. Fletcher, P. C. *et al. Neuroimage* **2**, 195–200 (1995).
7. Kosslyn, S. M. & Thompson, W. L. *Psychol. Bull.* **129**, 723–746 (2003).
8. Kosslyn, S. M., Thompson, W. L., Kim, I. J. & Alpert, N. M. *Nature* **378**, 496–498 (1995).
9. Van Essen, D. C. *et al. J. Am. Med. Assoc.* **41**, 1359–1378 (2001).

Supplementary information accompanies this communication on Nature’s website.

Competing financial interests: declared none.

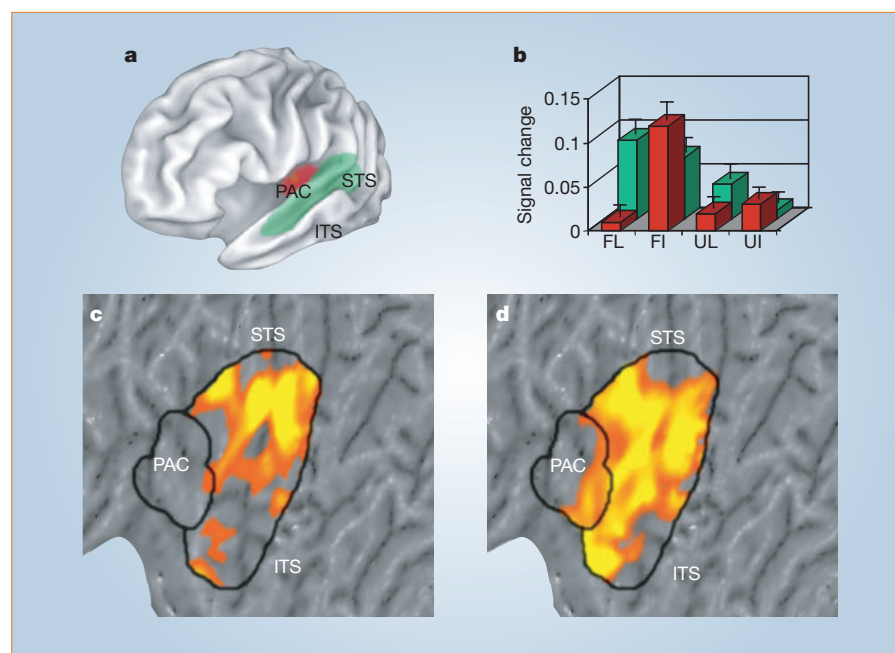


Figure 1 Auditory cortex activation during silent gaps in music. **a**, An inflated rendering of the left hemisphere⁹ illustrates primary auditory cortex (PAC; red) and auditory association cortex, also known as Brodmann’s area 22 (green). The superior temporal sulcus (STS) and inferior temporal sulcus (ITS) are indicated for reference. **b**, Signal change (arbitrary units) in PAC (red) and Brodmann’s area 22 (green) during gaps in familiar songs with lyrics (FL), familiar instrumentals (FI), unknown songs with lyrics (UL) and unknown instrumentals (UI). Error bars denote s.e.m. **c**, **d**, Difference in activity, which is greater for familiar songs, during silent gaps embedded in songs with (c) and without (d) lyrics, projected on to flattened views of the left temporal lobe. Dark-grey regions represent sulci; lighter grey regions denote gyri.

Imaging of Titan from the Cassini spacecraft

Carolyn C. Porco¹, Emily Baker¹, John Barbara², Kevin Beurle³, Andre Brahic⁴, Joseph A. Burns⁵, Sebastien Charnoz⁴, Nick Cooper³, Douglas D. Dawson⁶, Anthony D. Del Genio², Tilmann Denk⁷, Luke Dones⁸, Ulyana Dyudina⁹, Michael W. Evans³, Stephanie Fussner⁶, Bernd Giese¹⁰, Kevin Grazier¹¹, Paul Helfenstein⁵, Andrew P. Ingersoll⁹, Robert A. Jacobson¹¹, Torrence V. Johnson¹¹, Alfred McEwen⁶, Carl D. Murray³, Gerhard Neukum⁷, William M. Owen¹¹, Jason Perry⁶, Thomas Roatsch¹⁰, Joseph Spitale¹, Steven Squyres⁵, Peter Thomas⁵, Matthew Tiscareno⁵, Elizabeth P. Turtle⁶, Ashwin R. Vasavada¹¹, Joseph Veverka⁵, Roland Wagner¹⁰ & Robert West¹¹

¹Cassini Imaging Central Laboratory for Operations, Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, Colorado 80301, USA

²NASA Goddard Institute for Space Studies, 2880 Broadway, New York, New York 10025, USA

³Astronomy Unit, Queen Mary, University of London, London E1 4NS, UK

⁴C.E. de Saclay, Université Paris 7, L'Orme des Merisiers, 91191 Gif-sur-Yvette Cedex, France

⁵Department of Astronomy, Cornell University, Space Sciences Bldg, Ithaca, New York 14853, USA

⁶Department of Planetary Sciences, University of Arizona, 1629 E. University Blvd, Tucson, Arizona 85721, USA

⁷Institut für Geologische Wissenschaften, Freie Universität, 12249 Berlin, Germany

⁸Department of Space Sciences, Southwest Research Institute, 1050 Walnut Street, Suite 400, Boulder, Colorado 80302, USA

⁹Division of Geological and Planetary Sciences, California Institute of Technology, 150-21, Pasadena, California 91125, USA

¹⁰Institute of Planetary Research, German Aerospace Center, Rutherfordstrasse 2, 12489 Berlin, Germany

¹¹Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109, USA

Titan, the largest moon of Saturn, is the only satellite in the Solar System with a substantial atmosphere. The atmosphere is poorly understood and obscures the surface, leading to intense speculation about Titan's nature. Here we present observations of Titan from the imaging science experiment onboard the Cassini spacecraft that address some of these issues. The images reveal intricate surface albedo features that suggest aeolian, tectonic and fluvial processes; they also show a few circular features that could be impact structures. These observations imply that substantial surface modification has occurred over Titan's history. We have not directly detected liquids on the surface to date. Convective clouds are found to be common near the south pole, and the motion of mid-latitude clouds consistently indicates eastward winds, from which we infer that the troposphere is rotating faster than the surface. A detached haze at an altitude of 500 km is 150–200 km higher than that observed by Voyager, and more tenuous haze layers are also resolved.

In April 2004, as Cassini approached the saturnian system, the Imaging Science Subsystem¹ (ISS) began systematic observations of Titan with a three-month-long sequence, during which the pixel scale improved from ~200 km to 35 km. On 2 July 2004, shortly after entering Saturn orbit, Cassini had a distant (339,000 km) encounter with Titan (referred to as 'T0') during which images of the southern part of the sub-saturnian hemisphere and south-polar region were acquired at pixel scales of 2–3 km. The first close (1,200 km) targeted flyby of Titan ('TA') occurred on 26 October 2004, followed by another close pass with similar geometry ('TB') on 13 December 2004. Both achieved pixel scales of a few hundred metres. These new observations of Titan have major implications for the nature of the surface and atmosphere, and raise important new questions.

Titan's atmosphere consists primarily of nitrogen, with methane being the second-most abundant species (2–6%)^{2–4}. Photochemical reactions and escape of the lighter hydrogen produce a rich variety of complex hydrocarbons, which form a haze layer that obscures the surface at visible wavelengths and from which substantial quantities of organics may have precipitated onto the surface⁵. As a result, atmospheric methane has a short photochemical lifetime, suggesting that there is a source or reservoir of methane on or below Titan's surface. At Titan's surface temperature (~94 K) and pressure (1.5 bar), methane and ethane could be liquids, leading to suggestions of rainfall and lakes or seas of light hydrocarbons⁶. Earth-based telescopic observations had revealed bright and dark regions on the surface^{7–9} and ephemeral tropospheric clouds¹⁰. However, we knew very little about either Titan's surface or meteorology.

The new Cassini ISS observations reported here show intricate albedo markings as small as a kilometre on the surface, ~100 times finer than the best Earth-based images. In equatorial regions these markings exhibit east–west and other trends suggestive of aeolian transport and tectonic processes. Concentric or circular albedo patterns that could mark impact structures are rare, suggesting a relatively young surface modified by other geologic and atmospheric processes. Long, curvilinear features, which are primarily dark, but sometimes bright, may mark fluvial channels. However, no enhanced mirror-like (specular) reflections—which would indicate the presence of liquids on the surface—have been detected.

The images have also revealed important information about Titan's atmosphere. In addition to convective clouds in the south-polar (summer) region, we have detected mid-latitude tropospheric clouds. Some cloud features are elongated east–west; these are observed to occur preferentially at certain latitudes and, in the only case where height information is available, at lower altitude than the polar clouds. Tracking of cloud motions consistently yields eastward winds with speeds as high as 34 m s^{−1}, providing the first direct evidence of super-rotation in Titan's troposphere. Upper-atmospheric zonal banding is also apparent, being especially prominent in Titan's north-polar region. A well-defined haze layer is seen at an altitude of 500 km, significantly higher than the 300–350-km altitude layer seen during the Voyager 2 flyby in 1981 (ref. 11), and finer structure is resolved in higher-resolution images of the haze.

On the basis of the chemistry of its atmosphere, Titan is considered to be an important analogue to the early (pre-biotic) Earth¹². The Cassini ISS observations of Titan's surface suggest another strong similarity between these two bodies: the wide variety

of surface modification processes and their relative importance. Craters on Earth are relatively rare, being quickly modified by tectonism, erosion, burial and volcanism. Furthermore, on Earth all of these processes are continuous, so it is rare to be able to attribute the morphology of a particular region to a single process. Similarly, the diversity of the albedo patterns revealed on Titan's surface suggests a complex interaction between different processes acting concurrently to modify the surface.

Titan's surface

The surface of Titan was barely detected in Voyager visible-wave-length images acquired in 1981 (ref. 13). The Cassini ISS was designed¹ to take advantage of the improved surface visibility at particular infrared wavelengths where the optical depth of the atmospheric haze is lower and where methane gas is not strongly absorbing¹⁴. Observations acquired during approach were processed to produce a surface map with near-global coverage, except for the unilluminated north (winter) pole (Fig. 1). This map was similar to, but sharper than, the best Earth-based observations at 2.0–2.3 μm (refs 15, 16) where the haze optical depth is low, and it still provides the best-resolution view of the sub-saturnian, equatorial regions. Relatively low-albedo features with scales down to ~ 100 km, some with strikingly linear boundaries, are seen near the equator, whereas the middle latitudes appear more uniformly bright and the south-polar region is relatively dark. The large bright region centred at 10°S , 100°W has been named 'Xanadu Regio'; this is at present the only officially named feature on Titan.

Images of the southern part of the sub-saturnian hemisphere and south-polar region (2–3 km per pixel) were acquired in the T0 flyby (Fig. 2 and portions of Fig. 1). The low-latitude, large-scale features in these images agree well with the lower-resolution Saturn-approach and Earth-based imaging of Titan's surface, and significantly smaller-scale features are revealed. Several narrow (~ 10 – 20 km wide), dark features can be seen to extend for

hundreds of kilometres across the surface (Fig. 2b). In some cases these features are relatively straight, suggesting a tectonic influence on the albedo patterns, whereas in others they appear to wind across the surface as might be expected for liquid-carved channels. A broad (~ 300 km), dark feature runs northward between 330° and 0°W , and there are hints of a concentric pattern of dark, narrow, curvilinear features around the pole. Given the extent to which processes in Titan's atmosphere could affect the surface morphology, the differences between the surface patterns at equatorial and polar latitudes might reflect the cumulative effects of seasonal meteorological variations: for example, channels at high latitudes might be longer, deeper, wider and more extensive as a result of greater precipitation in polar regions (where clouds are currently most abundant).

The close (1,200 km) TA and TB flybys enabled imaging at lower phase angles ($\geq 14^\circ$ – 16°) than in T0 ($\geq 50^\circ$), and at much closer ranges. A mosaic of the regional-scale images (Fig. 3) reveals distinct patterns of relatively bright and dark areas. We interpret surface brightness variations, at all scales, as being due to the presence of surface materials with different albedos rather than topographic shading or shadows. Shadows probably cannot be seen from orbit at any phase angle at wavelengths shorter than $1\text{ }\mu\text{m}$. Even high-phase-angle images (Fig. 2) are likely to reveal only albedo markings for two reasons: (1) an icy satellite of this size is expected to have sufficiently low topographic relief¹⁷ that any shadows or shading would not be detectable at scales of 1–10 km, and (2) the atmospheric scattering must severely reduce the contrast between slope facets facing towards versus away from the direction of the Sun.

The cause of the albedo variations is still not certain, and there may be multiple explanations. It has been hypothesized that the darker regions are accumulations of hydrocarbons (liquid or solid) precipitated from the atmosphere and that the brighter regions are outcrops of water-ice bedrock with variable amounts of contamination^{18,19}. (Contaminants weathered from surface exposures of

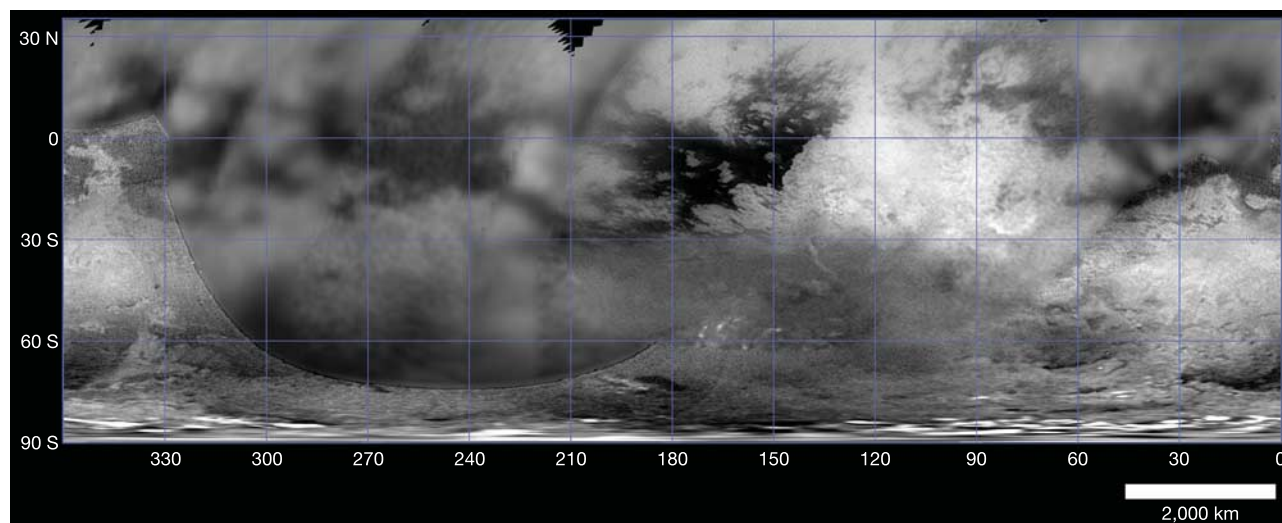


Figure 1 Albedo map of Titan's surface. Near-global, simple-cylindrical map of Titan's surface, assembled from Cassini ISS images taken during approach and the T0, TA and TB flybys. The pixel scales of individual images range from 88 to 2 km, and the corresponding resolutions of surface albedo features range from 176 to ~ 10 km. The very bright features near the south pole are clouds. High northern latitudes are not illuminated in this season. Unless otherwise noted, images presented here were acquired using the ISS filter that has proved to be most effective at peering through Titan's photochemical haze to its troposphere and surface, a narrow, continuum-band filter (CB3, 938 nm) centred in the best window in methane's absorption spectrum available to the ISS¹. Because the ISS has two filter wheels, we can also combine CB3 with an infrared polarizing filter, which further reduces the haze contribution at moderate ($\sim 90^\circ$) phase

angles, provided the spacecraft has the proper orientation with respect to Titan and the Sun^{1,44}. Images for all surface figures have been processed through the following steps: radiometric calibration; removal of obvious random noise; summation of two or three images to improve the signal-to-noise ratio; division by an image that only detects deep haze, preferably MT1 (619 nm), to normalize brightness variations with illumination and emission angle; improvement of camera-pointing information; re-projection of images to a common map projection; subtraction of 85% of a large-scale smoothed image to approximate atmospheric scattering of photons reflected from the surface; removal of data at very high emission and illumination angles; adjustment of brightness as a function of phase angle and use of polarizing filter; assembly into mosaics; and application of a seam-removal technique that preserves fine detail.

impure ice and transported by fluid flow could contribute to the composition of the darker regions.) This hypothesis and the ISS Titan observations (see below) are consistent with observations made by the Huygens Probe's Descent Imaging Spectral Radiometer (DISR) that indicate that bright areas along the probe's ground-track are relatively high standing and consist of contaminated water-ice, whereas dark areas are lower and have a higher proportion of non-ice material (M. Tomasko, personal communication). Radar measurements from Arecibo Observatory in Puerto Rico revealed a specular component at 75% (12 of 16) of the regions observed (globally distributed in longitude at $\sim 26^\circ$ S latitude), interpreted as indicative of dark, liquid hydrocarbons²⁰. However, there is no correlation between the places where specular signals were detected by Arecibo and the light and dark surface markings observed by the ISS (Fig. 4). There have been no reports to date of specular reflections at visible and near-infrared wavelengths, either by Cassini or Earth-based observers, although the optical depth of the atmosphere at these shorter wavelengths could limit the level at which such enhancement can be detected. That specular detections seem to be confined to radio wavelengths could indicate the presence of surface materials that are smooth at scales of centimetres but rough at scales of micrometres. Temporal variation in the presence of liquids on the surface is another possibility.

Analysis of data from Huygens will provide information about the presence of liquid hydrocarbons at the landing site, and further Cassini observations and analyses will help to put these local results into global and temporal context.

Another hypothesis to explain bright and dark regions is that ethane mist might overlie hydrocarbon liquids^{14,21}, perhaps marking exposures of liquid with bright diffuse features and obscuring specular reflections at optical wavelengths. We might expect such mists to vary in brightness and distribution over time, affected by wind, but comparison of TA and TB images revealed no apparent changes over the seven weeks between these encounters, except for the discrete clouds discussed below.

The albedo variations reveal only a few distinctly circular or concentric patterns that could indicate the presence of large impact structures. A dark annulus (outer edge $\sim 300 \pm 20$ km) surrounding a bright centre (Fig. 5) could delimit a large impact structure, with the former perhaps filled in part with dark materials in relatively low, confined areas by mass wasting or fluid flow²². Impacts by comets should produce a few hundred craters larger than 20 km in diameter per billion years²³. Furthermore, most of the ejecta from a near-catastrophic breakup of Hyperion was predicted to have struck Titan²⁴, as would ejecta from other impact events in the saturnian system. We have identified only five prominent

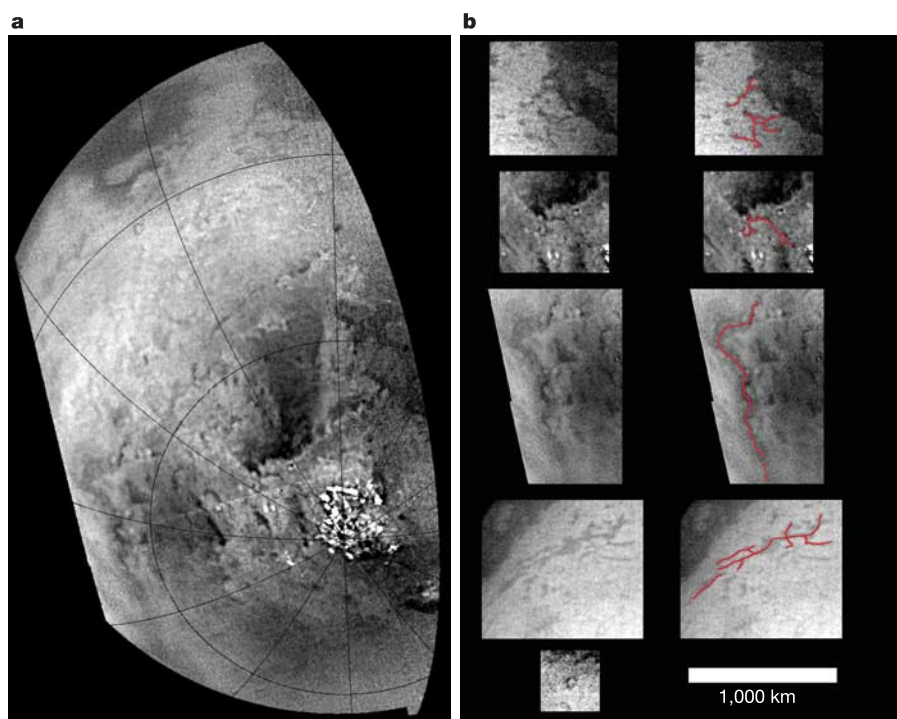


Figure 2 Images of Titan's south polar region. These images were acquired during Cassini's distant T0 encounter on 2 July 2004 and have pixel scales of ~ 2 km. **a**, A mosaic of Titan's south pole shown in an orthographic map projection. Lines of latitude and longitude are spaced at 30° intervals; the vertical meridian is 330° W. The small bright features near the south pole are clouds; the processing (see Fig. 1 legend) exaggerates their appearance. (The bright narrow line above the polar clouds and extending to the left is a seam in the mosaic.) **b**, Pairs of enlargements from **a** and other T0 images: original images (left) and versions in which some of the narrow, dark, curvilinear and rectilinear surface features, which may be examples of channels and/or tectonic structures, have been traced (right). The longest features (third and fourth pairs from the top) extend for more than 1,000 km across the surface and are as narrow as 10–20 km. At the bottom left, a single frame shows a small dark circular feature, which could be an impact crater. Although the pixel scales of the T0 images were ~ 2 km, no features smaller than ~ 10 km are resolved in the images, as expected¹. Most of the

photons detected by the ISS through the CB3 filter have been scattered by the atmospheric haze before ever reaching the surface, and ~ 60 – 90% (depending on phase angle and haze properties) of the light that is reflected off the surface from the nadir point (0° emission angle) is scattered by atmospheric haze, spreading the signal over regions up to ~ 100 -km wide. The ~ 10 – 40% surface reflection that passes through the atmosphere without scattering (or perhaps via strong forward scattering and relatively little smearing) returns a weak signal, and the modulation transfer function of the Cassini ISS at the Nyquist frequency reduces the contrast over about two pixels by an additional factor of 0.15 (ref. 1). As a result, with reasonable estimates of surface contrast, a single image with an overall signal-to-noise ratio of ~ 200 :1 cannot reveal surface features on scales smaller than about five pixels. The effects of the haze on surface visibility can also be seen by comparison of images showing the same region viewed at different emission angles: at high emission angles ($>45^\circ$) the surface contrast, and thus effective resolution, is significantly lower than when viewed at low emission angles (for example, Figs 2a and 4).

circular or concentric albedo features larger than 20 km in diameter (Figs 2b, 3, 5, 6a) over the ~30% of Titan imaged at better than 2 km per pixel. In particular, there are three bright circular features 30–50 km in diameter within the dark region west of Xanadu (Fig. 6a), which represents ~5% of the surface area of Titan. Given the lack of topographic information, craters might best be detected by the ISS as albedo features within dark regions where raised crater rims, presumably ice bedrock, would project above the dark infilling material. On the basis of the expected cratering rate²³, if these features are impact structures and we have detected all of the large craters in this region, the age of this part of Titan's surface (that is, the ice bedrock down to a depth of a few kilometres) is only ~130–300 Myr. If large impact structures are this rare on Titan, it suggests that the rates of surface modification processes such as viscous relaxation, tectonic resurfacing, erosion, deposition, and perhaps icy volcanism, have been sufficient to eliminate kilometre-scale relief over timescales of the order of 10^8 years. In contrast, surface ages for Ganymede and Callisto, which are expected to have levels of radiogenic heating comparable to Titan, generally exceed 10^9 years (ref. 23). One possible explanation for a geologically younger surface is that Titan is richer in ammonia, which lowers the melting point of water-ice and makes the resulting liquid neutrally buoyant²⁵, thereby allowing more extensive endogenic resurfacing at comparable levels of heat flow.

The albedo patterns we do see on Titan's surface are quite varied, and include diffuse bright streaks, narrow dark lineaments (both straight and curved), bright and dark patches with curved or

straight edges and often angular boundaries, and dark and bright spots (Figs 3 and 6). It is difficult to identify the processes responsible for many of these patterns, but likely candidates include aeolian and fluvial processes (perhaps sapping), icy volcanism, and tectonics. In many cases a combination of these processes is probably responsible for the features we observe, for example, tectonic structures modified by fluvial activity.

The prominent albedo boundary on the western side of Xanadu varies along its length (Fig. 3). In some places it is quite sharp and exhibits linear and angular shapes (Fig. 6a), whose form and orientation (at this resolution) are consistent with the dark material having embayed the bright terrain, an interpretation in agreement with the DISR findings that the bright regions near the Huygens landing site are higher than the dark regions (M. Tomasko, personal communication).

Streaks are ubiquitous in the equatorial region observed in TA and TB, generally trending west to east (Fig. 3). In addition, many bright patches within the large dark region have sharp western margins and diffuse, elongated eastern margins (Fig. 6b, c), also suggesting eastward flow. Finally, some of the bright patches have streamlined shapes consistent with movement of a fluid from west to east (Figs 3 and 6b). Given the large scale and extent of these features, the atmosphere seems the most likely fluid to be responsible for the patterns. The well-developed eastward pattern is not consistent with a tidal source, because tidal winds should reverse direction twice per Titan day²⁶. Instead, the preferred orientation suggests that aeolian features are controlled in some way by the

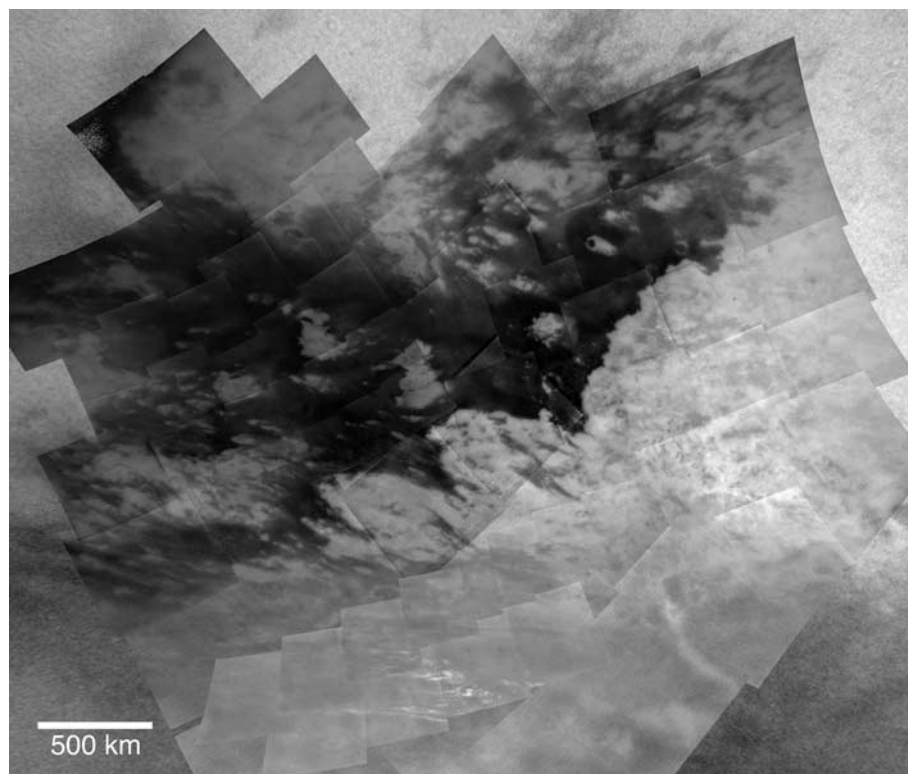


Figure 3 Mosaic of the equatorial region of Titan's anti-saturnian hemisphere. The background is a 3×3 mosaic acquired during TA at a scale of ~2 km per pixel (Fig. 4), with closer-range images (0.4–1.2 km per pixel, which means that the approximate scale of features that are resolved is 2–6 km) acquired during TA and TB covering most of the region. This is a simple-cylindrical map projection. Apart from the obvious clouds (described below), which are observed to move and change over timescales of several hours, we are confident that the local brightness variations in CB3 images represent surface features, because: (1) they have well-defined boundaries and have not been observed to change over time; and (2) the larger-scale features agree well with long-term,

Earth-based observations and those of the Cassini Visual and Infrared Mapping Spectrometer (VIMS)⁴⁵, acquired at longer wavelengths where atmospheric opacity is lower. The top-of-the-atmosphere normalized brightness I/F , where I is the intensity corrected to normal illumination and πF is the incident solar flux, varies from 0.26 to 0.39 in the CB3 images without the polarizer filter, but much of this brightness is due to haze. Because the haze is spatially uniform on the scale of the brightness variations, the ratio of the brightest to the darkest surface features is at least $0.39/0.26 = 1.5$. For the brightest surface features I/F is at least 0.13 ($= 0.39 - 0.26$) brighter than the dark material.

observed tropospheric super-rotation (see atmospheric discussion below). Eastward-biased gusts associated with turbulent eddy motions may be the most effective at entraining particles from the surface. Another possibility is wind streaks created as particles settle out of the atmosphere, although accumulation rates are relatively low²⁷.

Tectonic processes are also likely candidates for shaping Titan's surface: straight and angular lineaments are pervasive at global, regional and local scales (Figs 1, 3 and 6d, respectively). In the bright material on the anti-saturnian hemisphere, prominent lineaments often trend $\sim N 50^\circ W$ and $\sim N 80^\circ E$, like conjugate faults (Figs 3 and 6d). One interpretation of these features is large-scale tectonic modification of bright, bedrock material, with accumulations of darker material in low-lying areas enhancing the morphologic patterns (Fig. 6d). Another possibility is that transport of liquids within the ice bedrock results in sapping, which enlarges and darkens tectonic zones of weakness (Fig. 6e), similar to examples seen at closer range by DISR (M. Tomasko, personal communication). However, not all bright features may represent exposures of bedrock. Another hypothesis is that, in places, a thin brittle crust overlies relatively easily deformed dark material (perhaps organic materials precipitated from the atmosphere, although not

necessarily in the liquid state²⁸), and has been pulled apart or disrupted to expose dark regions (for example, Fig. 6f).

Continued monitoring of Titan may be essential for addressing questions about the causes of the albedo patterns observed on the surface (for example, whether streaks change direction over time, whether bright–dark boundaries change or move over time, whether dark linear features become more or less prominent in different seasons). We see no evidence for changes during the ~ 50 days between the TA and TB flybys, but changes may yet be seen over longer timescales and would help to constrain the processes responsible.

Titan's clouds and atmospheric dynamics

Methane clouds in Titan's troposphere have been observed in Earth-based images¹⁵. Before that, they were inferred from variability in near-infrared spectra¹⁰, and had been anticipated on the basis of estimated gaseous methane abundances and lapse rates. Cloud detections in ISS images are limited by the relatively short time intervals available for observations and the infrequent occurrence of clouds in most locations on Titan.

Three different types of clouds have been observed thus far. The most common are fields of small-scale bright clouds near the south pole (Fig. 7a–d), which were present during approach, T0 and TA, but not during TB. Their morphology and evolution is suggestive of cumulus convection. A movie showing the temporal evolution of the polar clouds during the approach to TA can be viewed at <http://ciclops.lpl.arizona.edu/view.php?id=537> (global view) or at <http://ciclops.lpl.arizona.edu/view.php?id=514> (polar region only). This behaviour is expected given the concentration of solar heating¹⁵ and rising motion at the summer pole^{29,30}. Simultaneous images in different filters (Fig. 8) show that the cloud feature is visible in a weak methane band (MT1, 619 nm) and even faintly in a somewhat stronger methane band (MT2, 727 nm). This observation

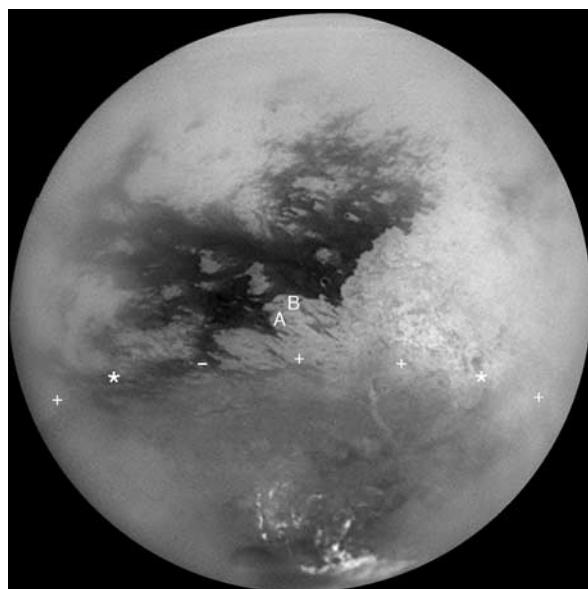


Figure 4 Locations of specular point observations on Titan's anti-saturnian hemisphere. Full-disk 3×3 mosaic of images acquired during TA at a pixel scale of ~ 2 km. The locations of specular points observed by ISS at 938 nm ('A' and 'B' for TA and TB, respectively) and from Arecibo at 13 cm (asterisk for sites of strong specular return and root-mean-square slope less than 1.2; minus sign for no specular component; plus sign for intermediate values of RMS slope) (ref. 20 and D. B. Campbell, unpublished work) are indicated. The TA and TB points are both near bright–dark boundaries, where a specular return could probably be seen from either unit, if there were optically smooth surfaces: for an optical depth of two and a liquid surface roughened by winds of 2 m s^{-1} , the region over which specular reflections would enhance the signal by at least 10% is roughly 200 km across or twice the size of the symbols A and B; for smoother surfaces (lower wind speeds) and/or lower optical depths the area with $>10\%$ enhancement can be several hundred kilometres across. Based on the root-mean-square slopes for the two strongest specular returns observed by Arecibo, the echo power comes from regions less than ~ 100 km across (centred on the sub-Earth point), which is slightly larger than the asterisks. However, owing to the motion of the sub-Earth point during the observation, the extent of the region that is the source of the specular detection represented by the asterisk on the right is known to extend at least 150 km in longitude²⁰. For the intermediate values of root-mean-square slope represented by plus signs, the source areas are 250–350 km across, which is 3–4 times the symbol sizes. Signal fluctuations observed at some locations implied structure on scales less than 10 km (ref. 20).

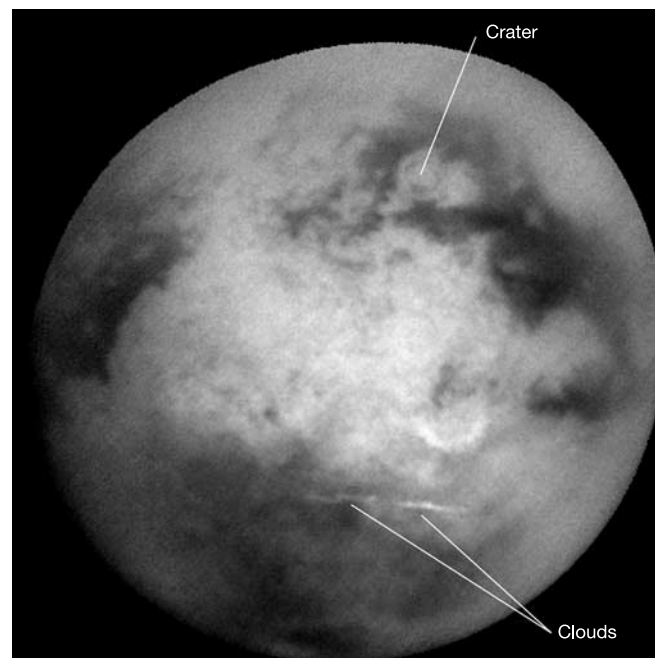


Figure 5 Full-disk image of Titan showing Xanadu Regio. Image acquired on 10 December 2004 during approach to the TB encounter with a pixel scale of 10.4 km. A concentric albedo pattern, perhaps an impact structure, can be seen within the dark region northeast of Xanadu Regio (bright region in centre of disk). The outer edge of the dark annulus has a diameter of 300 ± 20 km. Flow-like patterns southwest of this concentric feature could indicate sites of icy volcanism. Two bright, linear clouds are present just to the south of Xanadu.

suggests that the polar cloud tops are at high tropospheric altitude. Earth-based inferences suggest cloud tops as high as ~ 25 km (ref. 15).

A second, rarer cloud class is discrete mid-latitude clouds^{31,32} (Fig. 7e–g). These clouds tend to be small and faint and persist for only a few hours. We have thus far tracked cloud motions in 12 discrete mid-latitude and polar clouds using a digital tracking technique³³. Prograde (eastward) winds are measured for ten of the 12 clouds, the other two being motionless over the few-hour time interval (Fig. 9). The fastest wind speed recorded to date is $34 \pm 13 \text{ m s}^{-1}$ at 38°S latitude.

These wind vectors are the first direct evidence for Titan tropospheric super-rotation (atmosphere rotating faster than the surface), which had been inferred from Voyager temperature gradients³⁴ and predicted by models^{29,30,35,36}. Prograde winds can be produced by displacing rings of air from the equator to higher latitudes and conserving angular momentum, for example, in the near-tropopause branch of a Hadley cell. The cloud motions we

measure are probably from lower levels where winds are slower, but two of our wind vectors are faster than an angular-momentum-conserving profile (Fig. 9, dashed curve). A Hadley cell cannot produce such wind speeds if the rings of air are motionless at the equator, suggesting that eddy momentum transport is important in maintaining super-rotation on Titan³⁵.

Uniform angular momentum in the troposphere would require an unrealistically large meridional temperature gradient³⁷. Uniform atmospheric angular velocity relative to the surface (wind varying as cosine latitude) is more consistent with the small meridional surface temperature contrast inferred by Voyager³⁴ and more like the low-level super-rotation observed on Venus³⁷ and simulated by one Titan model³⁰. This would imply equatorial super-rotation speeds of tens of metres per second in the middle troposphere. Possible aeolian surface features implying eastward flow at low latitudes (Figs 3 and 6b) are also consistent with equatorial super-rotation.

The third class of cloud features consists of large-scale nearly zonal streaks (Figs 3 and 5). The 11 streak clouds observed to date

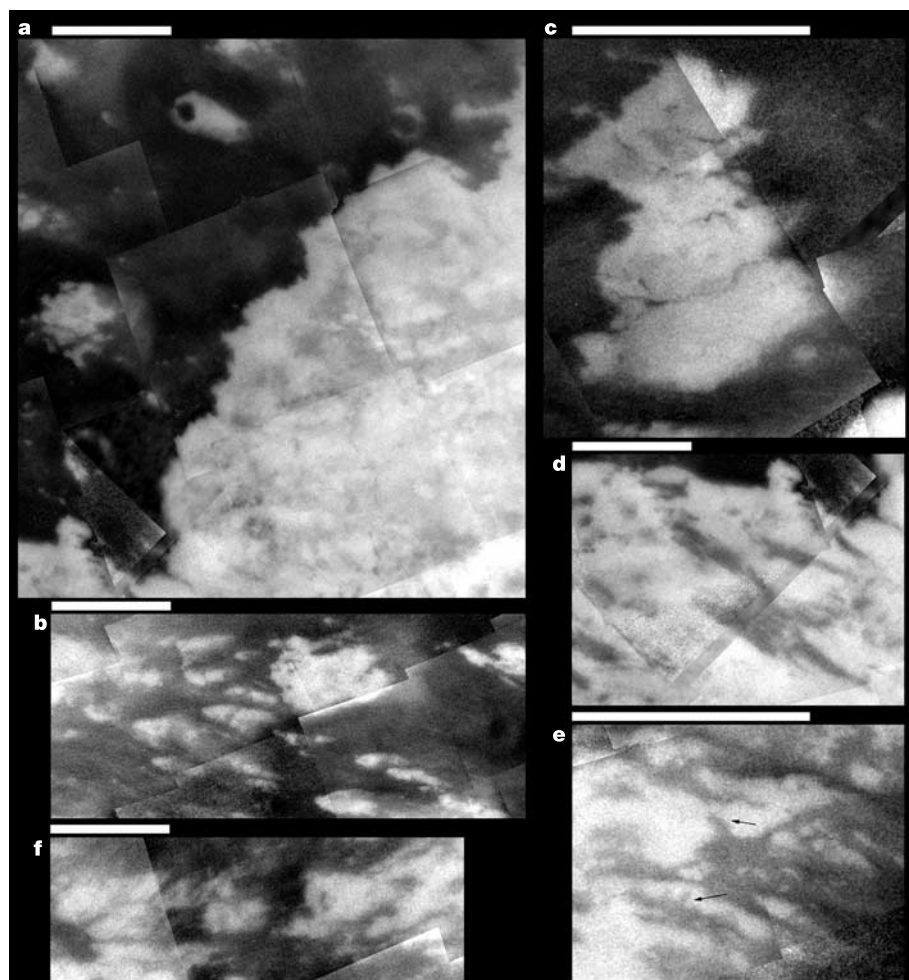


Figure 6 High-resolution views of Titan's surface. These enlarged views from the mosaic in Fig. 3 all have north to the top of the frame. White bars above each image are 200 km long. (Sharp lines are seams between individual images that have not been completely removed.) **a**, Prominent bright–dark boundary near the western edge of Xanadu exhibiting a sharp, angular contact between the two albedo units. Eastward-oriented, acute angles are common. The two bright, discontinuous circles, which could be impact structures, within the dark material near the top of the image are approximately 30 km in diameter and another near the lower left of the image is ~ 50 km in diameter. **b**, Streamlined bright features consistent with fluid motion from west to east. The Huygens landing site is in the upper left corner of this image. The region is in the centre left of Fig. 3. **c**, Bright feature

(near the centre of Fig. 3) surrounded by dark material, illustrating the contrast between the sharp western boundary and more diffuse eastern boundary. Several narrow (~ 2 km), dark lanes cross the bright area and the western bright–dark boundary exhibits acute angles as in **a**. **d**, Dark, linear lanes in this region west of Xanadu suggest a tectonic influence on the surface morphology, perhaps enhanced by other processes. **e**, Albedo patterns in this region southeast of the Huygens landing site (note arrows) exhibit shapes suggestive of erosion by sapping. **f**, A region near the upper right of Fig. 3 showing bright patches that could be reconstructed as though a bright crust has been extended east–west over a dark substrate.

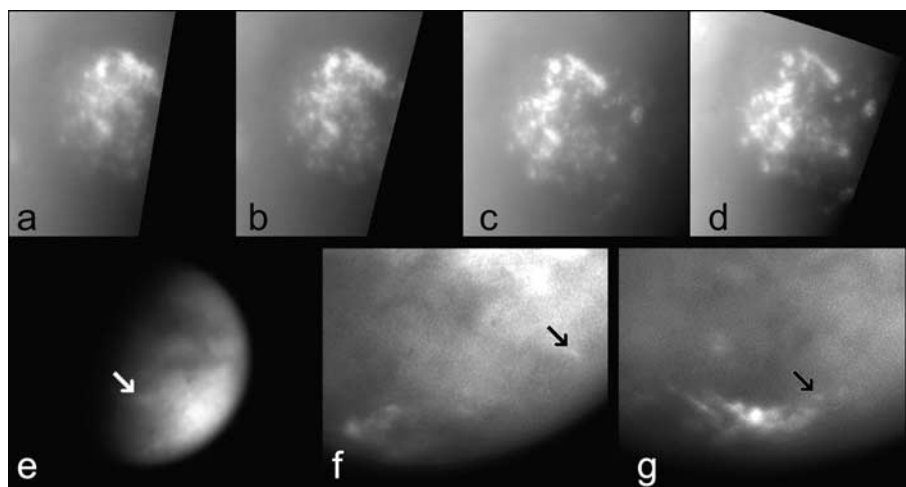


Figure 7 Tropospheric cloud features on Titan. **a–d**, A sequence of four methane continuum (IRP0-IR3, 928 nm) images showing the temporal evolution over the period 05:05–09:38 of the Titan south polar cloud field on 2 July 2004. **e–g**, Three examples of discrete mid-latitude clouds (arrows) for which motions have been tracked in CB3 images.

e, 38° S, 81° W (29 May 2004); this image was also viewed through an infrared polarizing filter. **f**, 43° S, 67° W (23 October 2004); **g**, 65° S, 110° W (25 October 2004). The clouds in **f** and **g** can also be seen briefly in the TA approach movies (see text for URLs).

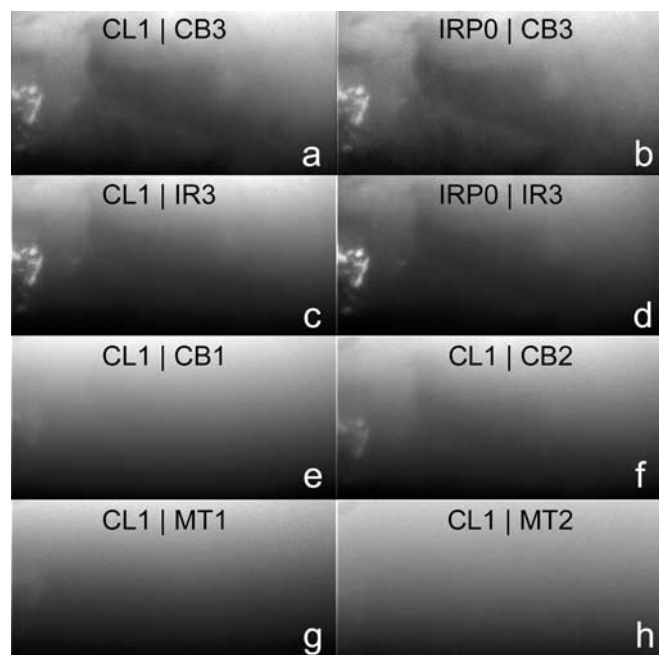


Figure 8 South polar clouds seen at different wavelengths. Near-simultaneous images of Titan's south polar regions acquired on 2 July 2004 in eight filter combinations, as indicated by the legend at the top of each image. The high-contrast cloud feature on the left of each image is located at the south pole. For each image, a spectral bandpass filter was combined with either a clear filter (CL1) or an infrared polarizing filter (IRP0), which helps reduce obscuration by the overlying stratospheric haze. Filters are as follows: Narrowband methane continuum images (CB3, 938 nm) without (**a**) and with (**b**) the polarizing filter. Broadband infrared (IR3, 928 nm) images, without (**c**) and with (**d**) the polarizing filter. **e**, **f**, Narrowband methane continuum images CB1 (635 and 603 nm; **e**) and CB2 (750 nm; **f**). **g**, **h**, Methane band images in weak (MT1, 619 nm; **g**) and moderate strength (MT2, 727 nm; **h**) bands. Both clouds and apparent surface features are visible in CB3 and IR3 and to a lesser extent in CB2 and CB1, for which obscuration by stratospheric haze is stronger. In MT1, all features except the polar clouds have disappeared. Gas absorption alone is sufficient to ensure that fewer than 1% of MT1 photons reach the surface, and haze scattering reduces this further. The cloud is barely visible in MT2, which has essentially no contribution from the surface.

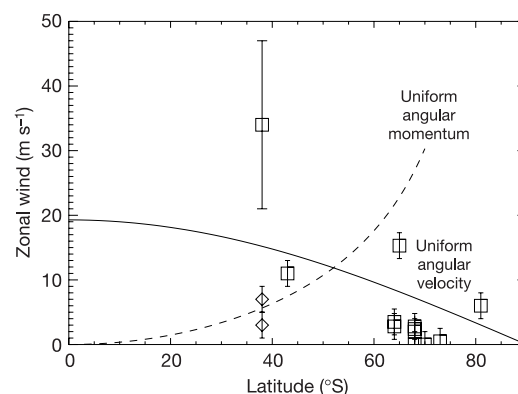


Figure 9 Cloud-tracked wind speeds versus latitude on Titan. The wind vectors probably refer to middle- and lower-troposphere altitudes. Squares indicate discrete clouds, and diamonds indicate streak clouds. The vertical lines are root-sum-square error bars for each wind speed. Latitude uncertainty is slightly smaller than the symbol widths. The scatter is due to real temporal variability, clouds at different altitudes with different mean winds, and tracking uncertainty. Wind speed uncertainty is the result of pixel resolution, navigation errors and cloud evolution. For clouds observed during the flybys, the pixel-resolution contribution is negligible (0.1 m s^{-1}). Navigation error is estimated by measuring displacements of nearby surface features; the worst-case error is 0.6 m s^{-1} . The largest error is due to cloud shape changes. We estimate a total wind speed uncertainty of 2 m s^{-1} based on changes in estimates for one feature at different times and differences between two tracking estimates. For the mid-latitude cloud feature seen on approach, pixel resolution is the dominant error source (9 m s^{-1}). Navigation error is much smaller, because digital tracking measures zero displacement of surface features in the same images. Assuming cloud evolution error to be less than one pixel, based on the correlation map, the total uncertainty is 13 m s^{-1} . The solid curve is a uniform angular velocity profile for the area-weighted mean angular velocity of all moving discrete features. It implies an equatorial super-rotation speed of $19 \pm 15 \text{ m s}^{-1}$, with most of the uncertainty probably due to real variation of wind speed with altitude. The dashed curve is a uniform angular momentum profile for a parcel with zero wind velocity at the equator. An angular-momentum-conserving profile consistent with our fastest wind vector implies a lower limit for the equatorial super-rotation of $22 \pm 10 \text{ m s}^{-1}$.

occur at preferred latitudes (six occurrences at 36–39°S, three at 67–72°S) and longitudes (all but two between 50–200°W), but examples exist at 14°S and 33°S as well. Cloud streaks are also seen in ground-based images at similar times³². The streak morphology and distribution invites three interpretations: (1) shearing and stretching by rapid upper-level super-rotating winds; (2) association with surface methane sources, for example, venting of methane gas from a subsurface reservoir through cracks in the surface; and (3) orographic waves generated by meridional flow over zonally oriented topography.

Each interpretation is problematic. Methane band images were available for one streak during TB, but it was not detected in MT2, indicating that it was at lower altitude than the polar clouds and thus less subject to shearing by strong winds. Two mid-latitude streaks were tracked at speeds of less than 10 m s^{-1} (Fig. 9), also implying tops at lower levels than other clouds and making it difficult to explain their large zonal extent unless these speeds are atypical. Some streaks curve slightly poleward from west to east, consistent with the expected sense of the summer lower troposphere meridional flow. Thus, generation from a surface source seems more likely, especially since streaks occur in preferred regions. If so, the

images do not yet identify the source. Medium-resolution images of a streak location during TB taken at a time when the streak was absent (TA) revealed no distinctive surface features, but cracks or topography finer than the $\sim 1\text{-km}$ effective ISS surface resolution cannot be ruled out.

Titan's stratospheric haze

Our early observations of Titan were made almost exclusively with filters at long or short wavelengths to optimize our chances of seeing the surface and tropospheric clouds or contrasts in the stratosphere, which would provide a basis for tracking motions. Images of the southern hemisphere with the UV3 filter (338 nm) showed no contrast, unlike images taken in 2003 using a similar filter on the Hubble Space Telescope³⁸ (HST) that do show a darker south polar hood. However, the HST images show that polar hood contrast decreased recently. The absence of contrast in Cassini images from April–July 2004 is consistent with that trend.

In UV3 images a faint thin haze layer encircles the denser stratospheric haze (Fig. 10a). Radial profiles of normalized intensity for this image (Fig. 11) suggest that the thin haze is similar to a 'detached' haze layer seen in Voyager images¹¹, but the altitude of the

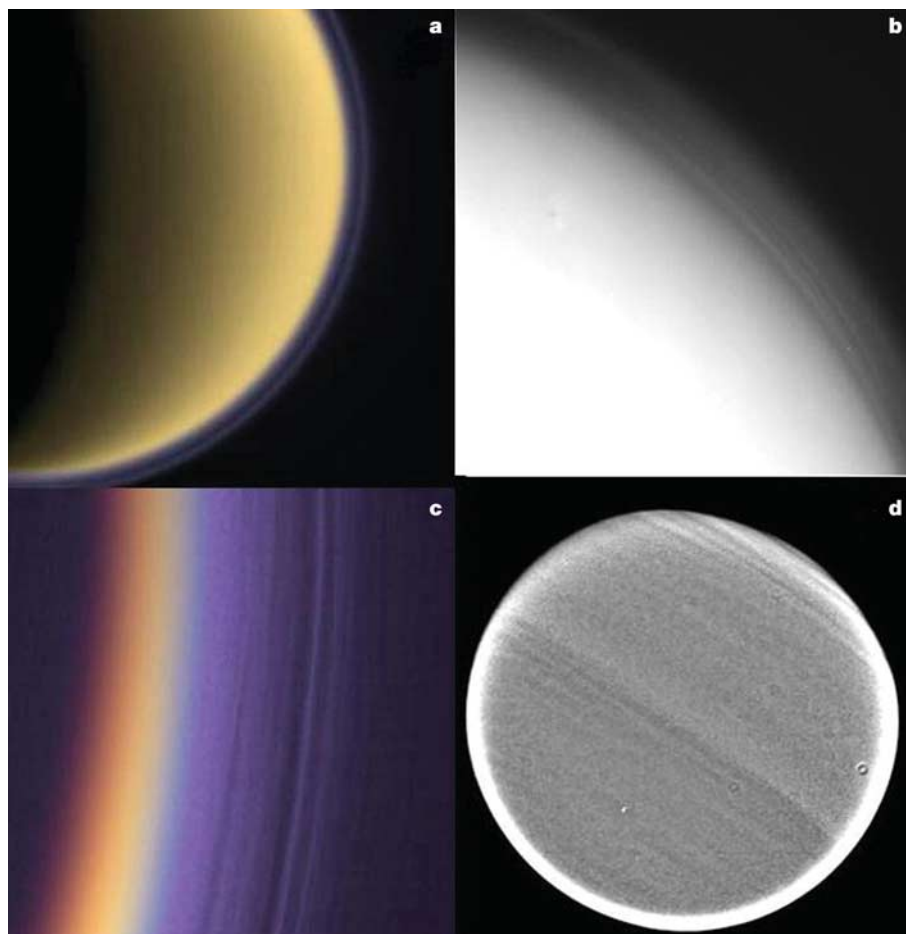


Figure 10 Images of Titan's stratospheric haze layers. **a**, Titan's southern polar region seen in the UV3 (338 nm) filter. The image is false-coloured to approximate what the human eye might see were our vision to extend into the ultraviolet: The globe of Titan retains the pale orange hue in the visible, and both the main atmospheric haze and thin detached layer are given their natural purple colour. The haze layers are brightened for visibility. The image was taken on 3 July 2004 at a Sun–Titan–spacecraft angle (or 'phase angle') of 114°. The pixel scale is 4.7 km. **b**, Titan's spin axis is tipped ~ 45 degrees to the vertical in this UV3 image of the northern polar hood. Multiple layers are seen between about 200 and 500 km altitude. **c**, A composite of a Cassini UV3 image of Titan's nightside limb, and BL1 (451 nm), GRN (568 nm), and RED (650 nm) filters taken within 4 min of

each other on 13 December 2004. They have been combined to look like true colour and stretched in brightness. The pixel scale is 0.7 km and the phase angle is 162°. The limb shown is at $\sim 10^\circ$ S latitude. Although this is a nightside view, with only a thin crescent receiving direct sunlight, the haze layers are bright from light scattered through the atmosphere. About 12 distinct haze layers can be seen extending several hundred kilometres above the surface. **d**, A methane filter (MT3, 890 nm) image shows faint horizontal banding at several latitudes, most prominently at northern high latitudes. The image was high-pass-filtered to enhance very weak contrasts at high spatial frequency. Consequently the limb appears very bright and there are circular features from dust on the optics—both artefacts amplified by the process.

haze observed by Cassini (near 500 km) is significantly higher than the detached haze seen by Voyager (300–350 km). One model suggests that the detached haze is produced by dynamics and haze sedimentation acting directly beneath the production altitude (which is specified to give the Voyager result)³⁰. The upward shift in haze altitude we observe thus suggests the possibility of seasonality in haze production or circulation strength, not captured by the existing models³⁹.

Images also reveal a multi-layer structure in the north polar hood region (Fig. 10b). The circle defined by the detached haze forms the top boundary of a set of arcs seen most readily in UV3 images. These arcs are not all mutually aligned, and two of them appear to merge. It is difficult to determine whether these are vertically distinct, individual haze layers or a manifestation of fewer layers with an altitude undulation that is accentuated in the near-terminator geometry of the north polar hood. Layering at lower latitudes is also seen in transmitted light looking back at high phase angle (Fig. 10c). These features might be produced by gravity waves, which have been detected previously at lower altitudes^{40,41}. Advection by the oscillating vertical component of wave motion in a haze of vertically varying concentration might make such waves visible. The vertical wavelength appears to decrease with altitude (Fig. 12).

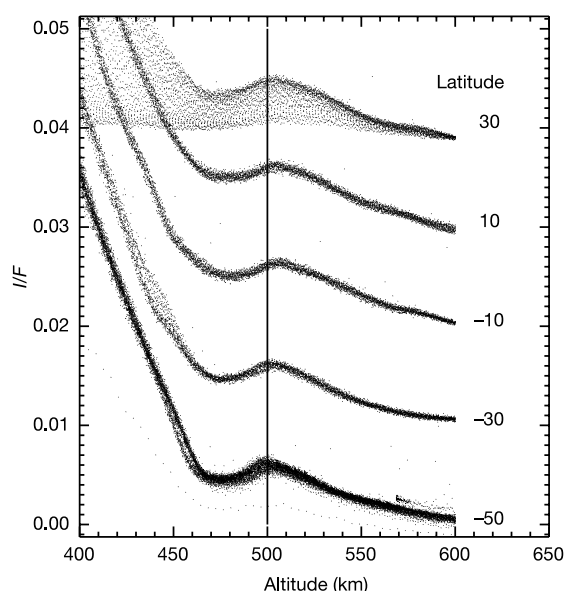


Figure 11 Intensity profiles of Titan's detached haze layer at varying latitudes. Reflected UV3 intensity (I/F) for five latitude bands versus altitude above Titan's surface. Latitude is defined as the latitude of the surface directly below the tangent point of the ray path that samples the stratosphere. Plots for latitudes north of -50° are offset by increments of 0.01 for clarity. Each set of points is selected from a bin 10° wide centred on the latitude shown. The points in the 30° bin include some near the terminator with low I/F values. Otherwise relative uncertainty can be gauged by the spread in the data for each profile, although there is a contribution from latitude variation and navigation uncertainty. Altitudes are sensitive to the position of Titan's centre in the image, which is uncertain to ± 3 pixels (about ± 15 km). There is also an I/F uncertainty of about ± 0.003 because the procedure for background subtraction was inexact. The images should also be deconvolved with the camera point spread function, which is expected to change the profiles slightly (the trough near 470 km will deepen and intensities will decrease faster above 500 km). However, these effects will not significantly change the altitude of the local I/F peak that defines the detached haze. The altitude of the detached haze was also measured from an image taken in late October when the phase angle was low and the detached layer was visible over most of the 360° of azimuth around the sub-spacecraft point. It is not visible as a detached layer in the polar hood region, but the circle fitted to the haze outside the polar hood also defines the boundary of the polar haze. This circle corresponds to an altitude of 507 km above Titan's surface.

This behaviour would be expected if the atmosphere became more thermodynamically stable with height, but these waves seem to be located in Titan's upper stratosphere/lower mesosphere where stability decreases with height. A more likely explanation for the decreasing wavelength, if real, is that the phase speed of the waves relative to the background zonal winds decreases upward as the super-rotation strengthens.

Higher-resolution images at high phase angle show many individual layers at low and middle latitudes, especially in the UV3 filter (Fig. 10c) but also in longer-wavelength filters. This and other images at high spatial resolution and high phase angle show undulations along the limb (apparent edge of the disk). These undulations are perhaps also a manifestation of a perturbing gravity wave. Structure is seen in the stratospheric haze against the disk with the MT3 (889N) filter, which samples the strong 890-nm methane absorption band. Weak contrasts with a strong zonal (axisymmetric) pattern are seen (Fig. 10d) and are most prominent at high northern latitudes. The contrasts are confined to the stratosphere at pressures < 150 mbar. Their strong zonal nature suggests a stratosphere dominated by zonal flow, as models predict^{29,30}. Eddies and meridional motion, if they exist at this level, are too weak to be detected.

Two global inversion layers above an altitude of 400 km have been inferred from analysis of stellar occultations⁴². These may well be related to the some of the layering we see. The inversions are at pressures of a few microbars where the sedimentation velocities of aerosols are high, so the new observations will be a challenge to aerosol microphysical models.

Discussion

Cassini's early imaging observations of Titan have begun to yield insights into the nature of its surface and atmosphere. Titan is a complex world, which appears to be influenced by tectonic, fluvial and atmospheric processes, in many ways similar to Earth, although the rates of such activity may be much slower on Titan.

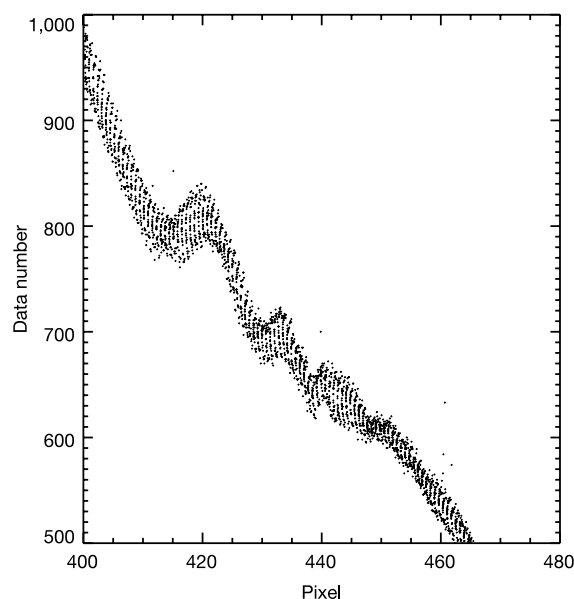


Figure 12 Multiple haze layers in Titan's north polar hood. Relative intensity (data number) in a radial direction from an altitude of approximately 200 to 500 km above the north pole of Titan. The layers seen in the image (Fig. 10b) appear here as local maxima in a downward-trending profile. The very weak local maxima and minima at sample numbers higher than 30 are visible as layers in the image. Altitude increases to the right. The relative uncertainty in the data number values can be judged from the vertical distribution of points. The data are taken from an unprocessed raw image.

The albedo patterns revealed on Titan's surface suggest that tectonics, viscous relaxation, aeolian, and perhaps fluvial processes have played significant roles in surface modification, although the ISS has yet to find direct evidence for surface liquids. Titan's current atmosphere inhibits the formation of impact craters of diameters less than ~20 km (ref. 43), so a lack of craters of these sizes would require either that the present atmosphere has persisted over a substantial part of Titan's history, or that the rate of surface modification has (at least recently) been rapid enough to remove or obscure smaller craters. However, large impact structures also appear to be rare, especially in comparison to the surfaces of other saturnian satellites, so other surface processes must remove kilometre-scale relief over timescales of 10^8 years. Furthermore, the widespread presence of linear features in equatorial regions implies that the long-term rates of tectonism must be comparable to, or greater than, those of erosion and deposition. Although spatial coverage by ISS is still somewhat limited, the overall style of the observed albedo patterns seems to vary with latitude, which may reflect differences in the cumulative effects of atmospheric processes that influence the surface. The analysis of Huygens data will provide important detailed information to aid our interpretations, and with continued Cassini observations we will build up a global view of Titan from which it will be possible to assess the various surface-modification processes at work and their relative importance. We will also be able to investigate the interactions between the surface and the atmosphere and look for changes as spring approaches in Titan's northern hemisphere.

Cassini images have also documented super-rotating winds in Titan's troposphere at mid-latitudes, verifying the predictions of general circulation models^{29,30,35,36}. Clouds and wind information at low latitudes is lacking at present, presumably because upwelling is concentrated at high latitudes in summer. If so, clouds might begin to occur at lower latitudes as Titan approaches the vernal equinox³², allowing equatorial super-rotation to be measured directly and compared with that observed by Huygens, and the possible role of precipitation in modifying the surface to be evaluated. Determining the vertical profile of the winds we measure will require quantifying cloud top altitudes, using radiative models constrained by methane abundance and stratospheric haze properties derived from Huygens data.

The altitude of Titan's detached stratospheric haze needs to be monitored as Titan approaches the equinox to determine whether the Voyager–Cassini difference is seasonal. The current results already provide a challenge to photochemical models to explain the factors that control the peak haze-production altitude and the processes that remove haze particles from the stratosphere. Such models are needed to estimate the downward flux of organics that may supply dark material to Titan's surface. □

Received 11 January; accepted 7 February 2005; doi:10.1038/nature03436.

1. Porco, C. C. *et al.* Cassini imaging science: Instrument characteristics and anticipated scientific investigations at Saturn. *Space Sci. Rev.* **115**, 363–497 (2004).
2. McKay, C. P. *et al.* Temperature lapse rate and methane in Titan's troposphere. *Icarus* **129**, 498–505 (1997).
3. Samuelson, R. E., Nath, N. R. & Borysow, A. Gaseous abundances and methane supersaturation in Titan's troposphere. *Planet. Space Sci.* **45**, 959–980 (1997).
4. Lunine, J. I. & Soderblom, L. A. Cassini-Huygens investigations of satellite surfaces and interiors. *Space Sci. Rev.* **104**, 191–208 (2002).
5. Khare, B. N., Sagan, C., Bandurski, E. L. & Nagy, B. Ultraviolet-photoproducted organic solids synthesized under simulated jovian conditions: Molecular analysis. *Science* **199**, 1199–1201 (1978).
6. Lunine, J. I., Stevenson, D. J. & Yung, Y. L. Ethane ocean on Titan. *Science* **222**, 1229–1230 (1983).
7. Smith, P. H., Lemmon, M. T. & Lorenz, R. D. Titan's surface, revealed by HST imaging. *Icarus* **119**, 336–349 (1996).

8. Coustenis, A. *et al.* Images of Titan at 1.3 and 1.6 μm with adaptive optics at the CFHT. *Icarus* **154**, 501–515 (2001).
9. Gibbard, S. G. *et al.* Speckle imaging of Titan at 2 microns: surface albedo, haze optical depth, and tropospheric clouds. *Icarus* **169**, 429–439 (2004).
10. Griffith, C. A. *et al.* Transient clouds in Titan's lower atmosphere. *Nature* **395**, 575–578 (1998).
11. Rages, K. & Pollack, J. B. Vertical distribution of scattering hazes in Titan's upper atmosphere. *Icarus* **55**, 50–62 (1983).
12. Raulin, F. & Owen, T. Organic chemistry and exobiology on Titan. *Space Sci. Rev.* **104**, 377–394 (2002).
13. Richardson, J., Lorenz, R. D. & McEwen, A. Titan's surface and rotation—New insights from Voyager images. *Icarus* **170**, 113–124 (2004).
14. Griffith, C. A., Owen, T. & Wagener, R. Titan's surface and troposphere, investigated with ground-based, near-infrared observations. *Icarus* **93**, 362–378 (1991).
15. Brown, M. E., Bouchez, A. H. & Griffith, C. A. Direct detection of tropospheric clouds near Titan's south pole. *Nature* **420**, 795–797 (2002).
16. Bouchez, A. *et al.* Keck Observatory Titan monitoring project. (<http://www2.keck.hawaii.edu/science/titan/>) (2005).
17. Perron, J. T. & de Pater, I. Dynamics of an ice continent on Titan. *Geophys. Res. Lett.* **31**, doi:10.1029/2004GL019802 (2004).
18. Lorenz, R. D. & Lunine, J. I. Titan's surface before Cassini. *Planet. Space Sci.* (in the press) (2005).
19. Griffith, C. A. *et al.* Evidence for exposure of water ice on Titan's surface. *Science* **300**, 628–630 (2003).
20. Campbell, D. B. *et al.* Radar evidence for liquid surfaces on Titan. *Science* **302**, 431–434 (2003).
21. Lunine, J. I. Does Titan have an ocean? A review of current understanding of Titan's surface. *Rev. Geophys.* **31**, 133–150 (1993).
22. Lorenz, R. D. Crater lakes on Titan: rings, horseshoes, and bullseyes. *Planet. Space Sci.* **42**, 1–4 (1994).
23. Zahnle, K., Schenk, P., Levison, H. & Dones, L. Cratering rates in the outer Solar System. *Icarus* **163**, 263–289 (2003).
24. Dobrovolskis, A. R. & Lissauer, J. J. Fate of ejecta from Hyperion. *Icarus* **169**, 462–473 (2004).
25. Croft, S. K., Lunine, J. I. & Kargel, J. Equation of state of ammonia-water liquid—Derivation and planetological applications. *Icarus* **73**, 279–293 (1988).
26. Tokano, T. & Neubauer, F. M. Tidal winds on Titan caused by Saturn. *Icarus* **158**, 499–515 (2002).
27. Strobel, D. F. Chemistry and evolution of Titan's atmosphere. *Planet. Space Sci.* **30**, 839–848 (1982).
28. Dimitrov, V. & Bar-Nun, A. Hardening of Titan's aerosols by their charging. *Icarus* **166**, 440–443 (2003).
29. Hourdin, F. *et al.* Numerical simulation of the general circulation of the atmosphere of Titan. *Icarus* **117**, 358–374 (1995).
30. Rannou, P. *et al.* A coupled dynamics-microphysics model of Titan's atmosphere. *Icarus* **170**, 443–462 (2004).
31. Lemmon, M. T. *et al.* HST observations of Titan, 1994–1997: New surface albedo maps and detection of large-scale, sub-tropical cloud activity. *Icarus* (submitted).
32. Roe, H. G. *et al.* Discovery of temperate latitude clouds on Titan. *Astrophys. J.* **618**, L49–L52 (2005).
33. Rossow, W. B., Del Genio, A. D. & Eichler, T. P. Cloud-tracked winds from Pioneer Venus OCPP images. *J. Atmos. Sci.* **47**, 2053–2084 (1990).
34. Flasar, F. M., Samuelson, R. E. & Conrath, B. J. Titan's atmosphere: Temperature and dynamics. *Nature* **292**, 693–698 (1981).
35. Del Genio, A. D., Zhou, W. & Eichler, T. P. Equatorial superrotation in a slowly rotating GCM: Implications for Titan and Venus. *Icarus* **101**, 1–17 (1993).
36. Del Genio, A. D. & Zhou, W. Simulations of superrotation on slowly rotating planets: Sensitivity to rotation and initial condition. *Icarus* **120**, 332–343 (1996).
37. Allison, M., Del Genio, A. D. & Zhou, W. Zero potential vorticity envelopes for the zonal-mean velocity of the Venus/Titan atmospheres. *J. Atmos. Sci.* **51**, 694–702 (1994).
38. Lemmon, M., Smith, P. & Lorenz, R. The evolution of Titan's polar hood near summer solstice. 35th COSPAR Scientific Assembly (Paris, 18–25 July 2004) abstr. COSPAR04-A-04048 (2004).
39. Hutzell, W. T. *et al.* Simulations of Titan's brightness by a two-dimensional haze model. *Icarus* **119**, 112–129 (1995).
40. Hinson, D. P. & Tyler, G. L. Internal gravity waves in Titan's atmosphere observed by Voyager radio occultation. *Icarus* **54**, 337–352 (1983).
41. Friedson, A. J. Gravity waves in Titan's atmosphere. *Icarus* **109**, 40–57 (1994).
42. Sicardy, B. *et al.* The structure of Titan's stratosphere from the 28 Sgr occultation. *Icarus* **142**, 357–390 (1999).
43. Korycansky, D. G. & Zahnle, K. J. Modeling crater populations on Venus and Titan. *Planet. Space Sci.* (in the press).
44. West, R. A. & Smith, P. H. Evidence for aggregate particles in the atmospheres of Titan and Jupiter. *Icarus* **90**, 330–333 (1991).
45. Sotin, C. *et al.* Cassini VIMS observations during the October 2004 flyby of Titan: Clues to the dynamics of Titan's surface. *Nature* (submitted).

Acknowledgements We acknowledge the many members of the imaging team who have assisted in the design of imaging sequences and camera commands and in other vital operational and image processing tasks, in particular N. Martin, E. Birath, J. Riley, B. Knowles, C. Clark, M. Belanger and D. Wilson. This work has been funded by NASA/JPL, the UK Particle Physics and Astronomy Research Council, the German Aerospace Center (DLR), and Université Paris VII Denis Diderot, CEA, AIM, France.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.C.P. (carolyn@ciclops.org).

Experimental one-way quantum computing

P. Walther¹, K. J. Resch¹, T. Rudolph², E. Schenck^{1*}, H. Weinfurter^{3,4}, V. Vedral^{1,5,6}, M. Aspelmeyer¹ & A. Zeilinger^{1,7}

¹Institute of Experimental Physics, University of Vienna, Boltzmanngasse 5, 1090 Vienna, Austria

²QOLS, Blackett Laboratory, Imperial College London, London SW7 2BW, UK

³Department of Physics, Ludwig Maximilians University, D-80799 Munich, Germany

⁴Max Planck Institute for Quantum Optics, D-85741 Garching, Germany

⁵The Erwin Schrödinger Institute for Mathematical Physics, Boltzmanngasse 9, 1090 Vienna, Austria

⁶The School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, UK

⁷IQOQI, Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Boltzmanngasse 3, 1090 Vienna, Austria

* Permanent address: Ecole normale supérieure, 45, rue d'Ulm, 75005 Paris, France

Standard quantum computation is based on sequences of unitary quantum logic gates that process qubits. The one-way quantum computer proposed by Raussendorf and Briegel is entirely different. It has changed our understanding of the requirements for quantum computation and more generally how we think about quantum physics. This new model requires qubits to be initialized in a highly entangled cluster state. From this point, the quantum computation proceeds by a sequence of single-qubit measurements with classical feedforward of their outcomes. Because of the essential role of measurement, a one-way quantum computer is irreversible. In the one-way quantum computer, the order and choices of measurements determine the algorithm computed. We have experimentally realized four-qubit cluster states encoded into the polarization state of four photons. We characterize the quantum state fully by implementing experimental four-qubit quantum state tomography. Using this cluster state, we demonstrate the feasibility of one-way quantum computing through a universal set of one- and two-qubit operations. Finally, our implementation of Grover's search algorithm demonstrates that one-way quantum computation is ideally suited for such tasks.

The quantum computer^{1,2} is a powerful application of the laws of quantum physics. Such a device will be far more efficient at factoring³ or database searches⁴ compared to its classical counterparts⁵. Considerable effort has been directed towards understanding the role of measurement and entanglement in quantum computation^{6–12}. A significant step forward in our understanding was the introduction of the one-way quantum computer^{13–17}, which highlights the importance of both measurement and entanglement in a striking way. In this model, all of the entanglement is provided in advance through a highly entangled multiparticle cluster state¹³. The quantum computation on the cluster state proceeds via local, single-qubit projective measurements with the outcomes potentially affecting those measurement settings that follow. It is a strength of the cluster state model that the intrinsic randomness of quantum measurement results creates specific types of errors that can be corrected through this classical feedforward. Most importantly, feedforward makes cluster state quantum computation deterministic. In the present proof-of-principle experiment, we perform measurements using fixed single-port polarizers, making our computations probabilistic. Different algorithms require only a different pattern of adapted single-qubit operations on a sufficiently large cluster state. As it is entirely based on single-particle measurements instead of unitary evolution, the computation is inherently not time-reversible—it is one-way. Most importantly, cluster state quantum computation is universal^{14,17} in that any quantum circuit can be implemented on a suitable cluster state. Open theoretical questions remain about the scalability under realistic noise conditions required for fault-tolerant one-way quantum computation. Although a threshold has been proved to exist¹⁸, it is unknown whether cluster state quantum computation will be more or less sensitive to noise than the standard model.

The one-way quantum computer does not perform quantum logic on the individual qubits of the cluster state. In order to describe the computational circuit, we need to distinguish between the physical qubits (in our case, the polarization state of photons),

which make up the cluster state and on which actual measurements are carried out, and encoded qubits, on which the computation is actually taking place. Owing to the specific entanglement of the cluster state, no individual physical qubit carries any information about an input state. Instead, each encoded qubit is written on the cluster state non-locally; that is, the information is carried by the correlations between the physical qubits. As the quantum computation proceeds, the encoded input qubits are processed in the imprinted circuit, whose output is finally transferred onto physical readout qubits. Interestingly, whereas the entanglement between the physical qubits in general decreases as a result of the measurement sequence, the entanglement between encoded qubits may increase.

A cluster state can be thought of as emerging from an array of equally prepared independent qubits, which then interact via controlled-phase ('CPhase') gates with their nearest (connected) neighbours. Specifically, a cluster state can be built up as follows: a large number of physical qubits are each prepared in the superposition state $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$, where $|0\rangle$ and $|1\rangle$ are the computational basis of the physical qubits. A CPhase operation $|j\rangle|k\rangle \rightarrow (-1)^{jk}|j\rangle|k\rangle$, with $(j,k \in 0,1)$, is then applied between pairs of neighbouring, connected qubits and effectively generates entanglement between them. The choice of which physical qubit neighbours are entangled by the CPhase operations, drawn as connecting 'bonds' (Fig. 1), determines the structure of the cluster state, which defines the basic type quantum circuit it can implement. This construction provides an intuitive understanding of the graphical representation of cluster states as connected arrays of physical qubits, in which each line corresponds to a previous nearest-neighbour interaction. We will later demonstrate how the highly entangled cluster states can be generated in a different way—directly from nonlinear optical processes.

Given a cluster state, two basic types of single-particle measurements suffice to operate the one-way quantum computer. Measurements in the computational basis $\{|0\rangle_j, |1\rangle_j\}$ have the effect of disentangling, that is, removing, the physical qubit j from the

cluster, leaving a smaller cluster state. Such operations can be used to modify the structure of the cluster and thus the imprinted circuit. The measurements that perform the actual quantum information processing are made in the basis $B_j(\alpha) = \{|\alpha\rangle_j, |-\alpha\rangle_j\}$, where $|\pm\alpha\rangle_j = (|0\rangle_j \pm e^{i\alpha}|1\rangle_j)/\sqrt{2}$ (α is a real number). The choice of measurement basis determines the single-qubit rotation, $R_z(\alpha) = \exp(-i\alpha\sigma_z/2)$, followed by a Hadamard operation, $H = (\sigma_x + \sigma_z)/\sqrt{2}$, on encoded qubits in the cluster ($\sigma_x, \sigma_y, \sigma_z$ being the usual Pauli matrices). Combinations of rotations about the z axis and Hadamard operations can implement $R_x(\alpha) = \exp(-i\alpha\sigma_x/2)$ rotations through the matrix identity $R_x(\alpha) = HR_z(\alpha)H$. General quantum logic operations can be carried out by the correct choice of $B_j(\alpha)$ on a sufficiently large cluster state. We define the outcome s_j of a measurement on the

physical qubit j as 0 if the measurement outcome is $|\alpha\rangle_j$, and as 1 if the outcome is $|-\alpha\rangle_j$. In those cases where the 0 outcome is found, the computation proceeds as desired. However, in those cases where the 1 outcome is found, a well-defined Pauli error is introduced. Feedforward, such that the output controls future measurement, compensates for these known errors.

For the implementations of single- and two-qubit quantum logic, we post-select only those cases where the 0 outcome is found. In these cases the computation proceeds error-free and requires no feedforward. In the final section, where we report the implementation of Grover's search algorithm, the feedforward determines the final, classical, measurement. There we measured all possible combinations of the measurement results individually, and applied the feedforward relation in such a way that the earlier

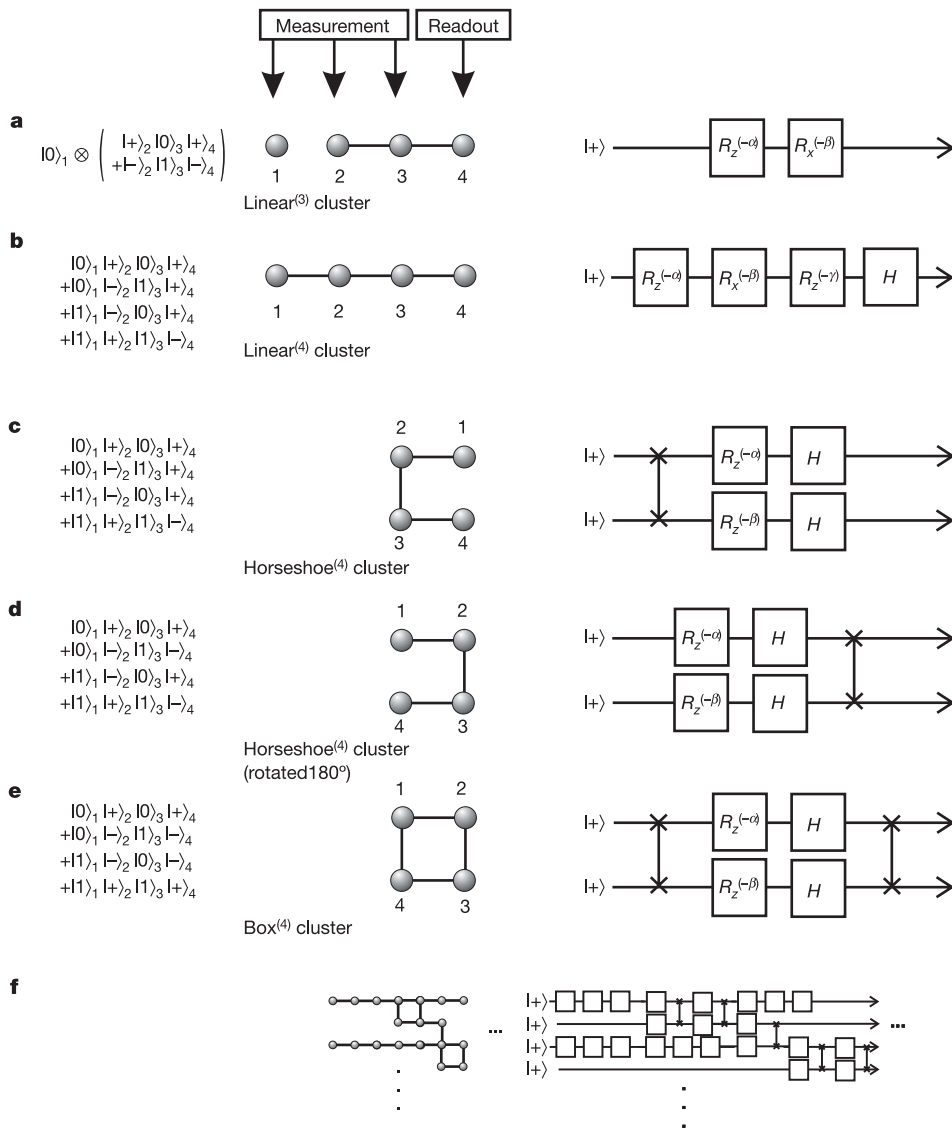


Figure 1 Few-qubit cluster states and the quantum circuits they implement. For each three- and four-qubit cluster, its quantum state and the computation carried out in the one-way quantum computer model are shown. For the case of the linear clusters $|\Phi_{\text{lin}3}\rangle$ (a) and $|\Phi_{\text{lin}4}\rangle$ (b), consecutive measurements on the qubits 1, 2 and 3 will perform a computation as a series of one-qubit rotation gates. The encoded input state undergoes single-qubit rotations with controllable angles, and the output is left on physical qubit 4. In contrast, the horseshoe clusters $|\Phi_{\text{c}4}\rangle$ (c) and $|\Phi_{\text{d}4}\rangle$ (d) and the box cluster $|\Phi_{\text{b}4}\rangle$ (e) form more complex circuits containing both single-qubit and two-qubit gates, both of which are necessary to form a universal set of logic gates for quantum computation. In

particular, measurements on two of the physical qubits (2 and 3 in the case of $|\Phi_{\text{c}4}\rangle$, and 1 and 4 in the case of $|\Phi_{\text{d}4}\rangle$ or $|\Phi_{\text{b}4}\rangle$) will perform the circuit defined by the particular cluster and transfer the logical output onto the remaining two physical qubits (1 and 4 in the case of $|\Phi_{\text{c}4}\rangle$, and 2 and 3 in the case of $|\Phi_{\text{d}4}\rangle$ or $|\Phi_{\text{b}4}\rangle$). When these cluster states are not a part of a larger cluster, the encoded input states are always $|\Psi_{\text{in}}\rangle = |+\rangle_{1E}$ for the one-qubit gates and $|\Psi_{\text{in}}\rangle = |+\rangle_{1E} |+\rangle_{2E}$ for the two-qubit gates. f, General input states can be prepared and processed through these operations when these clusters are subunits of larger clusters.

measurement outcomes define the meaning of the final ones.

Even a small cluster state suffices to demonstrate all the essential features of one-way quantum computing. Each of the three- and four-particle cluster states shown in Fig. 1 can implement the quantum circuit shown to its right that consists of a series of single- and two-qubit quantum gates. The computation proceeds via single-particle measurements carried out from the left side of the cluster to the right side, where the final readout takes place. The important feature of the quantum circuits is that the output of one circuit can be fed into the input of a subsequent one if their cluster states are bonded together by CPhase operations. Thus these small circuits, which form a universal set of logic gates, can be used as subunits for a fully functional quantum computer.

As an example, consider the four-particle box cluster state $|\Phi_{\square 4}\rangle$ on a two-dimensional lattice (Fig. 1e). The encoded input to the two-qubit quantum circuit is the product state $|\Psi_{\text{in}}\rangle = |+\rangle_{1E}|+\rangle_{2E}$, where the numerical subscript labels the qubit and the subscript E is used to distinguish encoded qubits from physical qubits. The circuit processes this pair of encoded qubits through a sequence beginning with a CPhase gate, followed by single-qubit rotations $R_z(-\alpha)$ and $R_z(-\beta)$ on encoded qubits 1 and 2, then a Hadamard operation, H , on both qubits, ending with a second CPhase operation. The values for α and β of the rotation gates are set by the choice of measurement bases $B_1(\alpha)$ and $B_4(\beta)$ on the physical qubits 1 and 4 respectively. The output of the quantum computation is transferred onto physical qubits 2 and 3. This kind of two-qubit quantum gate is essential for universal quantum computation, as it can generate entanglement between the encoded qubits.

On the other hand, by changing the geometry of the cluster state to the one-dimensional cluster $|\Phi_{\text{lin}4}\rangle$ (Fig. 1b), it now results in a different quantum circuit corresponding to a set of single-qubit rotations on one encoded qubit. Consecutive measurements $B_1(\alpha)$, $B_2(\beta)$ and $B_3(\gamma)$ on the physical qubits 1, 2 and 3 transform the input state, in our case $|\Psi_{\text{in}}\rangle = |+\rangle_{1E}$, to $|\Psi_{\text{out}}\rangle = HR_z(-\gamma)R_x(-\beta)R_z(-\alpha)|\Psi_{\text{in}}\rangle$ and store the output on qubit 4.

In order to demonstrate all the circuits shown in Fig. 1a–e, it is sufficient to produce first a linear cluster state of four qubits. The particular circuit implemented is then determined by the order of the measurements performed. Specifically, the one-dimensional linear structure (Fig. 1a and b) is implemented by sequentially measuring qubits 1, 2 and 3, with the final result then being available at qubit 4. The two-dimensional ‘horseshoe’ structures (Fig. 1c and d) are implemented by measuring either qubits 2 and 3 or 1 and 4, with the final result then being available at qubits 1 and 4 or 2 and 3, respectively. The four-qubit box cluster (Fig. 1e) can be obtained from the four-qubit linear cluster by Hadamard rotations and by swapping (that is, relabelling) the physical qubits 2 and 3. This will be described in more detail in the next section, and in the section ‘Two-qubit gates’.

The difficulty of one-way quantum computing lies with the cluster state preparation. Cluster states arise naturally in spin chains or spin lattices by way of nearest-neighbour Ising interaction¹³, a well-known interaction model in solid-state physics. The first proposals to achieve cluster states were based on analogues of dipole–dipole coupling between atoms in optical lattices^{15,19}. Although photon–photon interactions are negligible, recent proposals have nevertheless shown that optical systems may be well suited for implementing the cluster state model. These schemes use sequences of probabilistic quantum-logic gates^{8,20–23} based on linear optical elements to construct large photonic cluster states^{24,25}. These optical one-way quantum computation proposals are less demanding of resources than the comparable optical implementation in the standard model⁸.

In the present work, we have used nonlinear optics to directly produce four-photon cluster states. This method exploits a mode- and polarization-entangled four-photon state²⁶ produced in

pulsed-pump spontaneous parametric down-conversion (SPDC) (see Methods). We reconstructed the density matrix of the four-qubit cluster state using quantum state tomography, and studied the state’s entanglement properties relevant for quantum computation. We then implemented all of the quantum circuits shown in Fig. 1a–e, and demonstrated a two-qubit quantum search algorithm. In doing so, we have demonstrated the first universal set of gates and an important algorithm in a one-way quantum computer.

Creation and characterization of the cluster state

Our cluster state is produced experimentally using the mode- and polarization-entangled output of nonlinear spontaneous parametric down-conversion and linear optical elements, as described in detail in the Methods section. When four photons are emitted into the output modes of the polarizing beam-splitters 1, 2, 3 and 4, they are in the highly entangled cluster state

$$|\Phi_{\text{cluster}}\rangle = \frac{1}{2}(|H\rangle_1|H\rangle_2|H\rangle_3|H\rangle_4 + |H\rangle_1|H\rangle_2|V\rangle_3|V\rangle_4 + |V\rangle_1|V\rangle_2|H\rangle_3|H\rangle_4 - |V\rangle_1|V\rangle_2|V\rangle_3|V\rangle_4) \quad (1)$$

where $|H\rangle$ and $|V\rangle$ represent respectively horizontally and vertically polarized photon states, and the subscript labels the spatial mode. This state, $|\Phi_{\text{cluster}}\rangle$, is equivalent to the four-qubit linear cluster, $|\Phi_{\text{lin}4}\rangle$, and the horseshoe cluster states, $|\Phi_{\square 4}\rangle$ and $|\Phi_{\nabla 4}\rangle$ (Fig. 1) under the local unitary operation $H_1 \otimes I_2 \otimes I_3 \otimes H_4$ on the physical qubits, where H_i (I_i) is a Hadamard (identity) operation on qubit i . The four-qubit linear cluster can easily be reduced to a three-qubit linear cluster by measuring qubit 1 in the computational basis of the cluster and thus disentangling it from the rest. The state, $|\Phi_{\text{cluster}}\rangle$, can be converted to the box cluster state (Fig. 1e) by the local unitary operation $H_1 \otimes H_2 \otimes H_3 \otimes H_4$ and a swap (or relabelling) of qubits 2 and 3. Note that the four-qubit cluster state thus realized is also the smallest cluster state that represents a new kind of entanglement²⁷, while the two-qubit and three-qubit cluster states are locally equivalent to a maximally entangled Bell state and the three-qubit Greenberger–Horne–Zeilinger (GHZ) state, respectively.

State tomography

We have completely characterized our state via quantum-state tomography (which is a method for extracting the density matrix of a quantum state) from a discrete set of experimental measurements. Although quantum process and state tomography has been performed with up to three qubits^{28,29} this is, to our knowledge, the first time that a four-qubit density matrix has been determined from a complete, experimentally obtained density matrix. For a four-photon polarization state, like our cluster state, the full density matrix, ρ , is a 16×16 dimensional object that can be reconstructed by linear combinations of 256 linearly independent four-photon polarization projections. We performed each of these 256 measurements for 600 s using all combinations of $\{|H\rangle, |V\rangle, |+\rangle, |R\rangle\}$, where $|+\rangle = (|H\rangle + |V\rangle)/\sqrt{2}$ and $|R\rangle = (|H\rangle - i|V\rangle)/\sqrt{2}$. A maximum of 127 fourfold coincidence counts in 600 s were measured in the case of the setting $|H\rangle_1|H\rangle_2|V\rangle_3|V\rangle_4$. Instead of a direct linear combination of measurement results, which can lead to unphysical density matrices owing to experimental noise, we use a maximum-likelihood reconstruction technique^{30–32}. The resulting density matrix is shown in Fig. 2. The dominant diagonal elements represent the four components $|H\rangle_1|H\rangle_2|H\rangle_3|H\rangle_4$, $|H\rangle_1|H\rangle_2|V\rangle_3|V\rangle_4$, $|V\rangle_1|V\rangle_2|H\rangle_3|H\rangle_4$ and $|V\rangle_1|V\rangle_2|V\rangle_3|V\rangle_4$ as expected from equation (1), while off-diagonal elements indicate strong coherences between them. The negative coherences are due to the required π phase shift in the amplitude for the $|V\rangle_1|V\rangle_2|V\rangle_3|V\rangle_4$ term. Our reconstructed state is in good qualitative agreement with the target state, $|\Phi_{\text{cluster}}\rangle$. This can be quantified by the state fidelity $F = \langle \Phi_{\text{cluster}} | \rho | \Phi_{\text{cluster}} \rangle = (0.63 \pm 0.02)$, which is the state vector

overlap with the ideal state $|\Phi_{\text{cluster}}\rangle$. At present, no theoretical results exist concerning the fidelity requirements in cluster state quantum computation, so the full implications of the value that we found for the fidelity are unclear. Nevertheless, it has been proved that bi-separable four-qubit states cannot have a fidelity greater than 0.5 with our target cluster state. Our measured fidelity is clearly above this threshold, and therefore the observed fidelity is proof of the fact that the state contains the required four-particle entanglement³³ to a significant degree. Our cluster state is a coherent superposition of four different terms, each from a different physical origin (see Methods). The fidelity of our state is not perfect, because of partial distinguishability of these terms and because of imperfect phase stability in our set-up. We expect that achieving high fidelity with the target state will become rapidly more difficult as the size of the Hilbert space is increased. However, the acceptable amount of noise per qubit is independent of the size of the cluster, whereas it becomes exponentially small for GHZ states³⁴. Thus the problem of noise does not increase for larger clusters³⁴.

Entanglement properties

One-way cluster state quantum computation is based entirely on the entanglement properties of the physical qubits of the initial cluster state. For three-qubit states, there are only two classes of entanglement typified by the GHZ state³⁵, $|\text{GHZ}\rangle = (|H\rangle_1|H\rangle_2|H\rangle_3 + |V\rangle_1|V\rangle_2|V\rangle_3)/\sqrt{2}$, and by the W state^{36,37}, $|W\rangle = (|V\rangle_1|H\rangle_2|H\rangle_3 + |H\rangle_1|V\rangle_2|H\rangle_3 + |H\rangle_1|H\rangle_2|V\rangle_3)/\sqrt{3}$. As such, the three-qubit cluster state cannot represent a fundamentally different class of state and is, in fact, a GHZ state. On the other hand, our four-qubit cluster state cannot be converted via local unitary operations to either the four-qubit generalizations of the GHZ or the W state, but combines important characteristics of both.

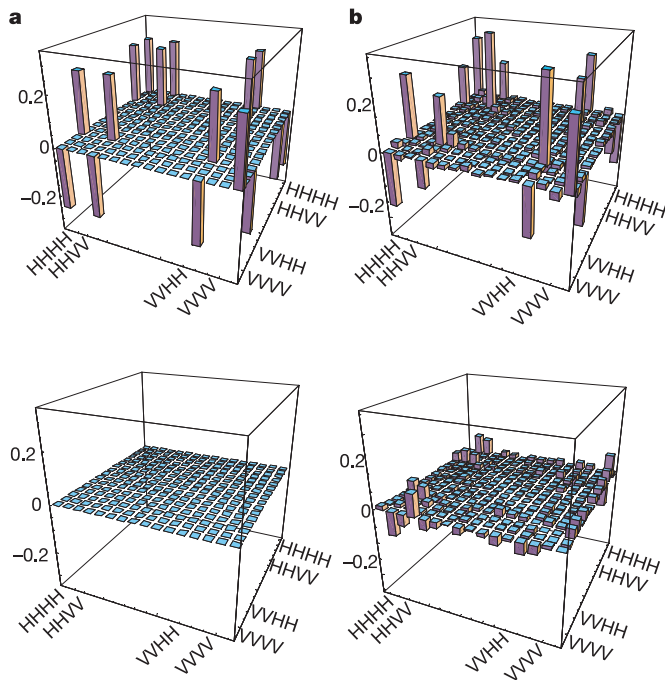


Figure 2 Density matrix of the four-qubit cluster state in the laboratory basis. Shown are the real (top) and imaginary (bottom) parts of the density matrix for the ideal case (a) and the reconstruction from the experimental four-photon tomography data (b). In both cases, there are four large diagonal components corresponding to HHHH, HHVV, VVHH and VVVV. The coherences between each of these diagonal elements show up as off-diagonal contributions, and are necessary for quantum entanglement. The real density matrix was reconstructed by way of a maximum likelihood method using four-photon coincidence rates obtained in 256 polarization projections.

As in a four-qubit GHZ state, we can make local measurements on one or two of the qubits and, with only classical communication, leave the remaining qubits in a three-qubit GHZ state or in a two-qubit Bell state, both of which can serve as an entanglement resource for quantum communication. For an explicit example, we reconstructed the density matrix of qubits 2, 3 and 4 by considering only the subset of our full tomographic measurements where qubit 1 was successfully measured in the state $|+\rangle_1$. Its fidelity compared to an ideal three-photon GHZ state was (0.60 ± 0.02) , which is above the local realism threshold³⁸ of 0.56.

Whereas the loss of one qubit in a GHZ state is already sufficient to completely disentangle the state, $N/2$ particles have to be removed from an N -particle cluster state in order to leave a separable state. The cluster states share this persistency of entanglement¹³ with the W states, which show an even stronger robustness against such decoherence. We demonstrate this property by considering the reduced density matrix of qubits 2, 3 and 4, $\rho_{2,3,4}^{\text{red}}$, after ignoring qubit 1—that is, after tracing out the first qubit. This trace has been implemented by summing the two subsets of our tomographic measurements in which the first qubit was measured in the state $|H\rangle_1$ or $|V\rangle_1$. To test for entanglement in the remaining three-qubit state, we used the entanglement-witness operator $W = \frac{1}{4}I^{\otimes 3} - \frac{1}{2}(|H\rangle_2\langle H|_2 \otimes |\Phi^+\rangle_{3,4}\langle\Phi^+|_{3,4} + |V\rangle_2\langle V|_2 \otimes |\Phi^-\rangle_{3,4}\langle\Phi^-|_{3,4})$, where $\text{Tr}(W\rho) \geq 0$ for all separable states, but is negative for some entangled states. Our state gives a value of (-0.115 ± 0.007) , which is negative by 16σ and thus proves that entanglement in qubits 2, 3 and 4 persists even after ‘loss’ of qubit 1. This remaining entanglement is between qubits 3 and 4. The loss of another particle destroys all entanglement and leaves a separable two-qubit state. We test this by calculating the eigenvalues of the partial transpose reduced density matrix of qubits 3 and 4, which are $\lambda_1=0.49$, $\lambda_2=0.45$, $\lambda_3=0.05$ and $\lambda_4=0.008$. These values are all positive, and thus fulfil the necessary and sufficient conditions for separability in two-particle systems^{39,40}.

A one-way quantum computer

Given a cluster state, quantum gates are implemented on the encoded qubits using only a series of single-qubit measurements and classical feedforward. In this section, we demonstrate the essentials of cluster-state quantum computation. In the present experiment, we created a cluster state and performed fixed polarization measurements, that is, projections onto the state $|+\alpha\rangle_j$ in the bases, $B_j(\alpha)$, which require no feedforward or subsequent

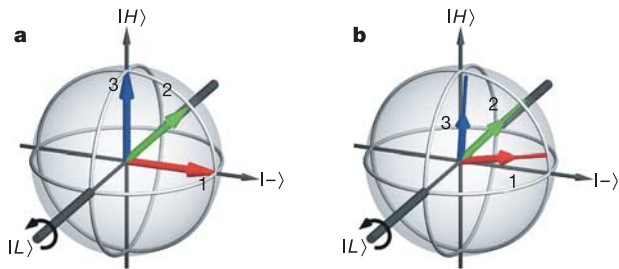


Figure 3 Output Bloch vectors from single qubit rotations using a three-qubit linear cluster $|\Phi_{\text{lin}3}\rangle$. The results of the ideal rotations (a) are compared with the results of the measured rotations (b) on the encoded input state $|+\rangle_{1E}$ for three different choices of measurement bases $B_2(\alpha)$. The output state is written in the laboratory basis. The measurement basis $B_3(\beta)$ for qubit 3 was fixed at $\beta = \pi/2$. The angle α is set to $\pi/2$, $\pi/4$ and 0 for the Bloch vectors labelled 1, 2 and 3 respectively. These different choices of α result in different rotations about the $|L\rangle$ axis. The sense of the rotation is shown for decreasing α . Our final states, extracted from measured single-qubit tomography, had fidelities of (0.86 ± 0.03) , (0.85 ± 0.04) and (0.83 ± 0.03) with respect to the ideal output states. Fidelities of outputs for other choices of angles and hence other rotations are shown in Supplementary Table 1b.

corrections given a successful measurement. This reduces the success rate of the computation by a factor of two for every measurement as compared to ideal, deterministic gate operations, but it certainly suffices as a proof of principle. An important challenge for future work is to implement the fast active switching and logic requirements for one-way quantum computation with feedforward. We have realized a universal set of quantum logic operations in the form of single-qubit rotations and non-trivial two-qubit gates. In the following sections we characterize the quality of each quantum computation by comparing the measured output state to the ideal using the state fidelity. Interesting avenues for study include full quantum gate characterization using quantum process tomography^{41,42} or related measures⁴³.

Single-qubit rotations

We start with the one-dimensional four-qubit linear cluster state, $|\Phi_{\text{lin}4}\rangle$, which implements an arbitrary single-qubit rotation gate (Fig. 1b). In particular, we perform $B_1(\alpha)$, $B_2(\beta)$ and $B_3(\gamma)$ on the physical qubits in the linear cluster basis. The parameters α , β and γ are sufficient to rotate the input qubit to anywhere on the Bloch sphere, or more generally to implement an arbitrary SU(2) single-qubit rotation. This computation rotates the encoded input qubit $|\Psi_{\text{in}}\rangle = |+\rangle_{\text{IE}}$ to the output state $|\Psi_{\text{out}}\rangle = HR_z(-\gamma)R_x(-\beta)R_z(-\alpha)|\Psi_{\text{in}}\rangle$, while the output of this computation is left in the quantum state of qubit 4. We finally characterize physical qubit 4 to verify the performance of the computation using single-qubit quantum state tomography (see Supplementary Table 1a).

The three-qubit linear cluster state $|\Phi_{\text{lin}3}\rangle$ is generated from $|\Phi_{\text{lin}4}\rangle$ by disentangling physical qubit 1 from the cluster. This is achieved by measuring physical qubit 1 in the computational basis for the linear cluster, that is, $\{|+\rangle_1, |-\rangle_2\}$ in the laboratory basis. We consider only those cases where we find the ‘+’ outcome. This

resulting cluster state implements the quantum circuit in Fig. 1a, which is a simpler single-qubit rotation gate with rotations determined by the measurements $B_2(\alpha)$ and $B_3(\beta)$ of the second and the third qubit. This rotates the encoded input qubit $|\Psi_{\text{in}}\rangle = |+\rangle_{\text{IE}}$ to the final state, $|\Psi_{\text{out}}\rangle = R_x(-\beta)R_z(-\alpha)|\Psi_{\text{in}}\rangle$, which is again left on physical qubit 4. The computation is directly implemented by performing single-particle measurements on qubits 1, 2 and 3. The single-qubit output stored on qubit 4 is completely characterized by single-qubit tomography. We compare this single output qubit with both the theoretically expected output and the predicted output from our reconstructed four-particle cluster state density matrix. Figure 3 shows the state of qubit 4 on the Bloch sphere in the laboratory basis in the ideal (left-hand side) and measured (right-hand side) case for three different measurement settings. These measurement settings were chosen to clearly show the effect of changing a single measurement basis. For the three state vectors shown as 1, 2 and 3, α was set to $\pi/2$, $\pi/4$ and 0, respectively, while β was fixed to $-\pi/2$. The state fidelities of these and other measurement settings compared to the ideal gate action are shown in Supplementary Table 1b.

Two-qubit gates

In order to perform universal quantum computation, non-trivial two-qubit quantum-logic gates⁴ are required in addition to single-qubit rotations. Well-known examples of such gates are the controlled-NOT (CNOT) or CPhase operations. The crucial trait of these gates is that they can change the entanglement between qubits. Gates of this type can be implemented with a two-dimensional cluster. In Fig. 1, we show two-dimensional cluster states—the horseshoe cluster states (Fig. 1c and d) and the box cluster state (Fig. 1e)—that satisfy this condition for two-qubit quantum logic. Both of their quantum circuits comprise single-qubit rotations and CPhase operations that can generate entanglement between two

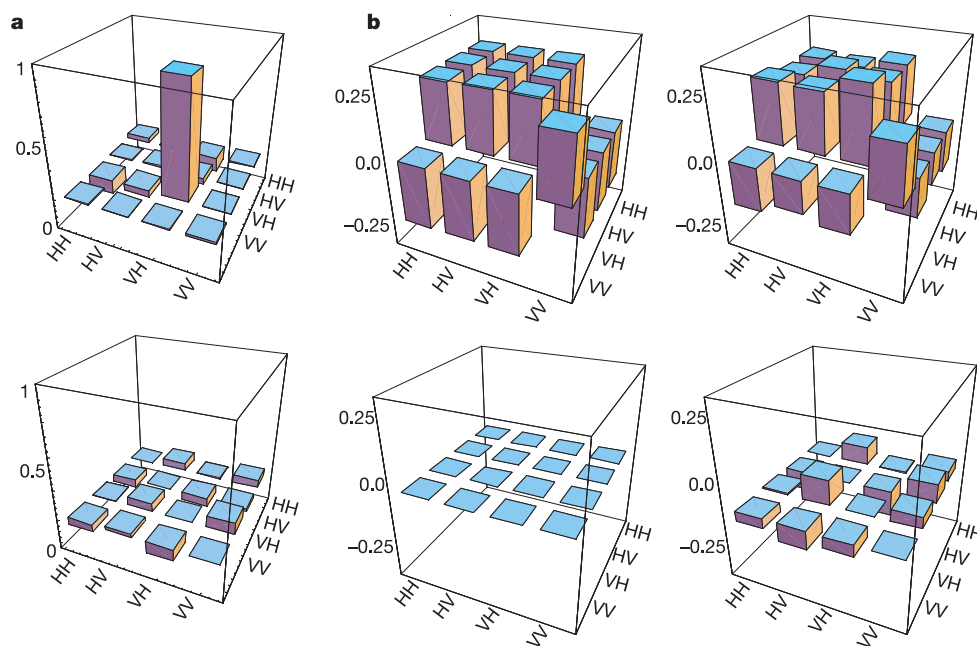


Figure 4 The output density matrices from two different two-qubit computations. Each density matrix is shown as two bar charts, with the upper bar chart depicting the real part of the matrix and the lower chart depicting the imaginary part. In case **a**, single-qubit measurements were made on qubits 1 and 4 in the box cluster state $|\Phi_{\text{box}4}\rangle$. The measurement settings were $B_1(\pi)$ and $B_4(0)$, which results in the expected output state $|V_2\rangle|H\rangle_3$ in the laboratory basis. The measured density matrix has (0.93 ± 0.01) fidelity with this state and no entanglement. In case **b**, single-qubit measurements in $B_2(0)$ and $B_3(0)$ were made on the horseshoe cluster state $|\Phi_{\text{C}4}\rangle$. In this case, the expected output

density matrix (left-hand side) and the experimentally measured density matrix (right-hand side) are both shown in the laboratory basis. The measured and expected density matrices are in good agreement and this is reflected in the fidelity (0.84 ± 0.03) . We extracted the tangle, a measure of entanglement, from the experimentally measured density matrix as $\tau = (0.65 \pm 0.11)$. This conclusively demonstrates that our cluster state quantum computer can generate the quantum entanglement necessary for universal quantum computation.

initially separable encoded qubits. Whether or not entanglement is generated depends on the specific circuit and the initial states. We will give a specific example of a two-qubit quantum computation that does not generate entanglement (in the box cluster) and a second computation that does generate entanglement (in the horseshoe cluster).

The box cluster transforms the two-qubit encoded input state according to

$$|\Psi_{\text{out}}\rangle = \text{CPhase}(H_1 \otimes H_2)[R_z(-\alpha) \otimes R_z(-\beta)]\text{CPhase}|\Psi_{\text{in}}\rangle \quad (2)$$

where α and β are determined by measurements $B_1(\alpha)$ and $B_4(\beta)$ on qubits 1 and 4, respectively. For the isolated box (or horseshoe) cluster state, the encoded qubit input state is the product state $|\Psi_{\text{in}}\rangle = |+\rangle_{1E}|+\rangle_{2E}$. Consider the case where photons 1 and 4 are measured to be $|V\rangle_1$ and $|H\rangle_4$ in the laboratory basis. For the box cluster, this corresponds to measuring the '0' outcome of $B_1(\pi)$ and $B_4(0)$ respectively on physical qubits 1 and 4. According to the box quantum circuit, this should perform the computation $|+\rangle_{1E}|+\rangle_{2E} \rightarrow |+\rangle_{1E}|-\rangle_{2E}$ with the resulting product state outcome left on qubits 2 and 3. The experimentally measured final state of qubits 2 and 3, in the laboratory basis, are shown as a two-qubit density matrix in Fig. 4a. The state has a single dominant diagonal element corresponding to the state $|V\rangle_2|H\rangle_3$ and no off-diagonal elements. Recall that the conversion from the box basis to the laboratory basis requires a swap of qubits 2 and 3 and the application of a Hadamard on each qubit. This converts $|+\rangle_{1E}|-\rangle_{2E}$ to $|V\rangle_1|H\rangle_{2E}$, in agreement with our measured density matrix. The fidelity of this measured density matrix of qubits 2 and 3 with the ideal output is (0.93 ± 0.01) . We can

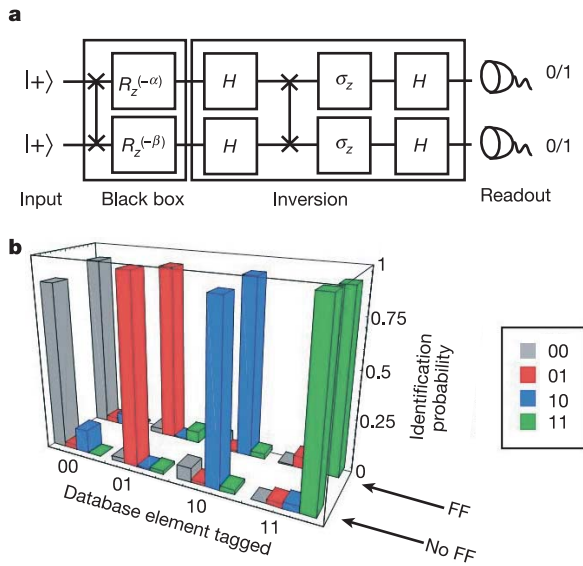


Figure 5 Grover's algorithm in a cluster state quantum computer. **a**, The quantum circuit implementing Grover's search algorithm for two qubits. The box cluster state implements the quantum circuit shown in Fig. 1e. These two circuits perform the equivalent computation when the readout measurements on physical qubits 2 and 3 in the box cluster are carried out in the bases $B_{2,3}(\pi)$. The rotations, and hence black-box function, are fixed by the measurement settings $B_1(\alpha)$ and $B_4(\beta)$ on physical qubits 1 and 4. The circuit implements one of the four black boxes in the search algorithm, and processes the output through an inversion-about-the-mean operation. The output, which is the final states of physical qubits 2 and 3, reveals which black box was applied. **b**, The experimentally measured outputs of this quantum computation. The data labelled 'no FF' show the computational outputs $\{s_2, s_3\}$ in those cases where the measurements in the black box found the '0' outcome. The data labelled 'FF' show the outputs for all individually measured outcomes from the black box to which the feedforward relation $\{s_2 \oplus s_4, s_3 \oplus s_1\}$ has been applied to the output results. The probability for successful identification of the function is approximately 90% in all cases.

quantify the entanglement in the output state using the tangle⁴⁴ defined as $\tau = [\text{Max}(0, \sqrt{\lambda_1} - \sqrt{\lambda_2} - \sqrt{\lambda_3} - \sqrt{\lambda_4})]^2$, where λ_i are the eigenvalues of the matrix $\rho \Sigma \rho^T \Sigma$ and $\Sigma = \sigma_y \otimes \sigma_y$, and are decreasingly ordered with λ_1 being the largest. The tangle can range from 0 for separable states to 1 for maximally entangled states. The tangle of our measured density matrix is (0.01 ± 0.01) , in agreement with the expected value of 0—no entanglement has been generated in this case. The fidelities of other quantum computations in the box cluster are shown in Supplementary Table 2a, including an example where entanglement is generated.

Encoded two-qubit operations are crucial for cluster state quantum computation as they can generate entanglement. The horseshoe cluster state of Fig. 1c performs the following quantum circuit:

$$|\Psi_{\text{out}}\rangle = (H_1 \otimes H_2)[R_z(-\alpha) \otimes R_z(-\beta)]\text{CPhase}|\Psi_{\text{in}}\rangle \quad (3)$$

For our input state $|\Psi_{\text{in}}\rangle = |+\rangle_{1E}|+\rangle_{2E}$, this circuit always generates maximal entanglement. Consider the case where photons 2 and 3 are both measured in the state $|+\rangle$. This measurement corresponds to the '0' outcome of $B_2(0)$ and $B_3(0)$ on physical qubits 2 and 3, and should perform the transformation $|+\rangle_{1E}|+\rangle_{2E} \rightarrow \frac{1}{\sqrt{2}}(|0\rangle_{1E}|+\rangle_{2E} + |1\rangle_{1E}|-\rangle_{2E})$, where the output is maximally entangled. The experimentally measured output density matrix of photons 2 and 3 is shown in Fig. 4b (right-hand side). For

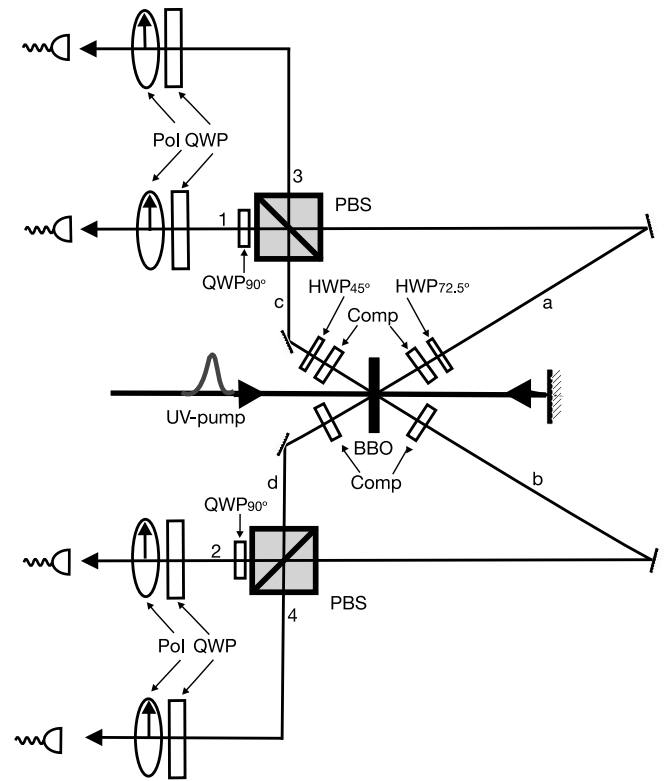


Figure 6 The experimental set-up to produce and measure cluster states. An ultraviolet laser pulse makes two passes through a nonlinear crystal (BBO), which is aligned to produce entangled photon pairs $|\Phi^-\rangle_{a,b}$ in the forward direction in modes a and b, and $|\Phi^+\rangle_{c,d}$ in the backward direction in modes c and d. Compensators (Comp), consisting of a half-wave plate (HWP) and an extra BBO crystal, are placed in each path to counter walk-off effects. Including the possibility of double-pair emission and the action of the polarizing beam-splitters (PBSs), the four components of the cluster state can be prepared. The incorrect phase on the HHVV amplitude can easily be changed by using a HWP in mode a. The amplitudes can be equalized by adjusting the relative coupling efficiency of those photon pairs produced in the backward pass as compared to the forward pass. Polarization measurements were carried out in modes 1 to 4 using quarter-wave plates (QWPs) and linear polarizers (Pol) followed by single-mode fibre-coupled single-photon counting detectors behind 3-nm interference filters.

comparison, the ideal output density matrix is also shown in Fig. 4b (left-hand side). The two density matrices are qualitatively very similar and, indeed, the state fidelity of our measured state with the ideal state is (0.84 ± 0.03) . The tangle of this output state is $\tau = (0.65 \pm 0.11)$, confirming the generation of entanglement between the logical qubits as a result of the quantum computation. Furthermore, this experimentally measured density matrix implies a maximum Clauser–Horne–Shimony–Holt (CHSH) Bell parameter⁴⁵ of $S = (2.47 \pm 0.08)$, which is well above the $S=2$ upper limit for local realistic theories. The fidelities of other quantum computations in the horseshoe cluster are shown in Supplementary Table 2b.

Grover's search algorithm

The excitement over quantum computation is based on just a few algorithms, the most well known being Shor's factorization algorithm³ and Grover's search algorithm⁶. The latter is extremely important, both from a fundamental standpoint, as it is provably more efficient than the best classical algorithm, and from a practical standpoint, because fast searching is central to solving difficult problems. Grover's algorithm has been implemented under the standard model^{46,47} both in NMR^{48,49} and in optical experiments⁴⁷. Here, we demonstrate a two-qubit implementation of Grover's fast quantum search using the cluster state model.

The goal of a search is to identify one out of N elements of a database. Formally, one could consider the database as a black box that labels one of N possible inputs leaving all others unchanged. The challenge, then, is to find that labelled state. The best classical strategy chooses the inputs randomly one by one and checks for the label; it takes, on average, about $N/2$ calculations to find the special input. Quantum parallelism in Grover's algorithm allows for all inputs to be processed simultaneously in superposition, while interference enhances the probability of finding the desired outcome in only $O(\sqrt{N})$ calculations of the function. In the case of a two-bit function ($N=4$), the difference is even more dramatic, as Grover's algorithm needs only one calculation of the function, whereas classically three evaluations are needed in the worst case, and 2.25 evaluations are needed on average.

The quantum circuit for the two-qubit Grover algorithm is shown in Fig. 5a. Two input qubits, 1 and 2, are prepared in the state $|+\rangle_1|+\rangle_2$. This is a superposition of all four computational basis states $|0\rangle|0\rangle$, $|0\rangle|1\rangle$, $|1\rangle|0\rangle$ and $|1\rangle|1\rangle$. As one of four possibilities, the black box could label the state $|0\rangle|1\rangle \rightarrow -|0\rangle|1\rangle$ by changing its sign and leaving all other states unchanged. Note that the change of sign is equivalent to a bit flip on an ancillary qubit. Any of the four possibilities can be set by proper choices of the rotation angles α and β . The output from the black box is processed by an operation that inverts the amplitudes for each computational state about the mean value. This process amplifies the labelled amplitude and reduces the rest. In the two-qubit case, theory predicts that after a single application of this inversion, the computer outputs the labelled state with probability 1.

We can compare this circuit for the two-qubit Grover algorithm to that implemented by the box cluster state in Fig. 1e. The Grover algorithm circuit contains extra fixed single-qubit operations (a σ_z followed by a Hadamard transformation, H) on each qubit before the readout in the computational basis. Measurement of a final physical qubit in the computational basis after a σ_z followed by H is equivalent to direct measurement of those qubits in the basis $B(\pi)$, that is, we can absorb these additional fixed single-qubit operations into the readout stage. The quantum circuit implemented by the box cluster can be seen as precisely that one required for Grover's algorithm provided that the final readout measurements are made in $B_{2,3}(\pi)$ on physical qubits 2 and 3.

The cluster state computation proceeds as follows. The encoded qubits begin in the state $|+\rangle_1|+\rangle_2$. Setting the measurement angles, α, β , to $\pi, \pi, 0, 0$ determines the black-box settings

00, 01, 10 and 11, respectively. In principle, these settings remain hidden. The result of a measurement of physical qubit i is s_i , which is 0 (1) for a measurement of $|+\alpha\rangle$ ($|-\alpha\rangle$). For the computation to proceed deterministically, the black box must provide the measurement outcomes for feedforward. The encoded qubits are transferred non-locally to the remaining physical qubits in the cluster. Remarkably, the inversion-about-the-mean process is already 'hard-wired' into the structure of the cluster state and is automatically implemented. The output of the computation, including feedforward (FF), are two bits $\{s_2 \oplus s_4, s_3 \oplus s_1\}$ identifying the black box.

In Fig. 5b, we show the experimental outcomes of the quantum computation. As in the previous computations, measurements were made using quarter-wave plates and polarizers. The 'no FF' data are the computational outputs obtained when the black-box outcomes, s_1 and s_4 , were 0, which requires no feedforward but reduces the success rate to 1/4. In addition, we have measured individually all possible correlations between the measurement results from the black box and the readout. This enables us to implement the simplest feedforward possible, where the earlier measurement determines the meaning of the final readout. Thus, when the black-box measurement outcomes are other than $s_1=0$ and $s_4=0$, it is necessary to reinterpret the readout via the bitwise addition shown above; the 'FF' row of data in Fig 5b shows the sum of the readouts interpreted in this way. In either case, the probability of the quantum computer determining the correct outcome is about 90%. These high-fidelity results (Fig. 5b) constitute, to our knowledge, the first demonstration of a quantum algorithm in a cluster state quantum computer.

Discussion

We have generated a four-qubit cluster state, and characterized that state and its entanglement features. With that cluster, and taking advantage of a curious equivalence of a number of cluster states, we have demonstrated a universal set of single-qubit and non-trivial two-qubit quantum logic gates in a one-way quantum computer. Our final realization of the Grover algorithm strongly underlines the basic simplicity of the cluster state approach. Given the various alternatives for their creation, such as linear optics, ion traps and optical lattices, and the recent advances in the preparation of multiparticle entangled states, cluster states are promising for inclusion in future implementations of quantum computation. The most important challenges for the optical approach presented here are (1) realization of cluster states on demand, (2) generating cluster states with more qubits and (3) implementation of fast feedforward where earlier measurement outcomes actually change the setting of a future measurement in real time. $\square$

Methods

Experimental preparation of cluster states

In our experiment, entangled photons are created by using type-II parametric down-conversion⁵⁰. A frequency-doubled laser pulse at 795 nm makes two passes through a β -barium borate (BBO) crystal, which emits highly entangled photons into the forward pair of modes a and b and the backward pair of modes c and d (Fig. 6). To counter the effect of birefringence in the BBO crystal, the polarization in each mode is rotated by 90° and the photons pass through compensation crystals that erase transverse and longitudinal walk-off. Final half-wave plates (HWP), one for each photon pair, and the tilt of the compensation crystals allow for the production of any of the four Bell states. The modes of the forward emitted pairs a and b and the modes of the backward emitted pairs c and d are coherently combined at polarizing beam-splitters (PBSs) by adjusting the position of the delay mirror for the ultraviolet pump. The preparation of the cluster state is based on the simultaneous emission of four photons. The construction of the set-up allows for four photon events to come from either two entangled pairs, one forward and one backward, or from double-pair emission into the modes a and b , and c and d ²⁶.

Proposed methods for producing cluster states are based on series of two-qubit gates, such as the CPhase or CNOT. In our case, the four-photon cluster state is generated directly from parametric down-conversion. Because of its intrinsically probabilistic nature, the down-conversion process becomes exponentially inefficient for generating larger cluster states. Our way of generating the cluster states furthermore exploits the properties of PBSs and uses post-selection. A PBS is an optical device that transmits horizontally polarized light and reflects vertically polarized light. Considering the two-photon case, where after the PBS in each mode one photon has to be detected, both

incoming photons must have the same polarization when they come from different input modes, or they must have orthogonal polarizations when entering along the same input mode. If the source produces simultaneously a $|\Phi^-\rangle_{a,b}$ state into the forward pair of modes, and backwards a $|\Phi^+\rangle_{c,d}$ state, only the state $|H\rangle_1|H\rangle_2|H\rangle_3|H\rangle_4 - |V\rangle_1|V\rangle_2|V\rangle_3|V\rangle_4$ results in a four-photon coincidence. However, there exists also the case where a four-photon emission into the two modes on either side results in a fourfold coincidence. The state must be in this case $-|H\rangle_1|H\rangle_2|V\rangle_3|V\rangle_4$ coming from the $|\Phi^-\rangle_{a,b}$ setting and $+|V\rangle_1|V\rangle_2|H\rangle_3|H\rangle_4$ coming from the $|\Phi^+\rangle_{c,d}$ setting. The final emerging state is a superposition of all four terms.

In order to produce a cluster state, the phase of the $-|H\rangle_1|H\rangle_2|V\rangle_3|V\rangle_4$ term has to be shifted by π . This can be done using a HWP in one mode, where a polarization rotation of an angle ϕ causes the state after the PBSs to evolve to $-\cos(2\phi)|H\rangle_1|H\rangle_2|V\rangle_3|V\rangle_4$. Thus any HWP rotation larger than 22.5° results in a sign flip. At the same time, the Bell state is converted to $\cos(\phi)|\Phi^-\rangle_{a,b} + \sin(\phi)|\Psi^+\rangle_{a,b}$, where the amount of the wanted $|\Phi^-\rangle_{a,b}$ state is decreased by a factor of $\cos(\phi)$. Owing to the properties of the PBS, only the amplitudes for the $|H\rangle_1|H\rangle_2|H\rangle_3|H\rangle_4 - |V\rangle_1|V\rangle_2|V\rangle_3|V\rangle_4$ terms are affected, while the $|\Psi^+\rangle_{1,2}$ Bell state does not contribute to any fourfold coincidence.

Note that after each PBS a quarter-wave plate (QWP) was placed to compensate for birefringence effects. For each measurement, the phase of the back-reflected pair or four-photon was kept fixed and verified for each measurement setting. Taking into account the emission rates of the source for the entangled pairs ($28,000\text{ s}^{-1}$ two-photon coincidences for the forward-emitted pair, and $18,000\text{ s}^{-1}$ coincidences for the backward-emitted pair), theoretical calculations show that a HWP rotation by 27.5° results in the maximally entangled cluster state of the form $|\Phi_{\text{cluster}}\rangle = \frac{1}{2}(|H\rangle_1|H\rangle_2|H\rangle_3|H\rangle_4 + |H\rangle_1|H\rangle_2|V\rangle_3|V\rangle_4 + |V\rangle_1|V\rangle_2|H\rangle_3|H\rangle_4 - |V\rangle_1|V\rangle_2|V\rangle_3|V\rangle_4)$. Thus, a HWP in mode a has been rotated to prepare this state. Fine-tuning has been done by short measurements to obtain approximately equal count rates for each component.

Received 7 October 2004; accepted 11 January 2005; doi:10.1038/nature03347.

1. Deutsch, D. & Ekert, E. Quantum computation. *Phys. World* **11**, 47–52 (1998).
2. Braunstein, S. L. & Lo, H.-K. (eds) Experimental proposals for quantum computation. *Fortschr. Phys.* **48** (special focus issue 9–11), 767–1138 (2000).
3. Shor, P. W. in *Proc. 35th Annu. Symp. Foundations of Computer Science* (ed. Goldwasser, S.) 124–134 (IEEE Computer Society Press, Los Alamitos, 1994).
4. Grover, L. K. Quantum mechanics helps in search for a needle in a haystack. *Phys. Rev. Lett.* **79**, 325–328 (1997).
5. Bennett, C. & DiVincenzo, D. Quantum information and computation. *Nature* **404**, 247–255 (2000).
6. Ekert, A. & Jozsa, R. Quantum algorithms: entanglement enhanced information processing. *Phil. Trans. R. Soc. Lond. A* **356**, 1769–1782 (1998).
7. Gottesman, D. & Chuang, I. L. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature* **402**, 390–393 (1999).
8. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409**, 46–52 (2001).
9. Linden, N. & Popescu, S. Good dynamics versus bad kinematics: Is entanglement needed for quantum computation? *Phys. Rev. Lett.* **87**, 047901 (2001).
10. Jozsa, R. & Linden, N. On the role of the entanglement in quantum computational speed-up. *Proc. R. Soc. Lond. A* **459**, 2011–2032 (2003).
11. Nielsen, M. A. Quantum computation by measurement and quantum memory. *Phys. Lett. A* **308**, 96–100 (2003).
12. Biam, E., Brassard, G., Kenigsberg, D. & Mor, T. Quantum computing without entanglement. *Theor. Comput. Sci.* **320**, 15–33 (2004).
13. Briegel, H. J. & Raussendorf, R. Persistent entanglement in arrays of interacting particles. *Phys. Rev. Lett.* **86**, 910–913 (2001).
14. Raussendorf, R. & Briegel, H. J. A one-way quantum computer. *Phys. Rev. Lett.* **86**, 5188–5191 (2001).
15. Raussendorf, R. & Briegel, H. J. Computational model underlying the one-way quantum computer. *Quant. Inform. Comput.* **2**, 344–386 (2002).
16. Raussendorf, R., Brown, D. E. & Briegel, H. J. The one-way quantum computer—a non-network model of quantum computation. *J. Mod. Opt.* **49**, 1299–1306 (2002).
17. Raussendorf, R., Brown, D. E. & Briegel, H. J. Measurement-based quantum computation on cluster states. *Phys. Rev. A* **68**, 022312 (2003).
18. Nielsen, M. & Dawson, C. M. Fault-tolerant quantum computation with cluster states. Preprint at (<http://arXiv.org/quant-ph/0405134>) (2004).
19. Mandel, O. *et al.* Controlled collisions for multiparticle entanglement of optically trapped ions. *Nature* **425**, 937–940 (2003).
20. O'Brien, J. L., Pryde, G. J., White, A. G., Ralph, T. C. & Branning, D. Demonstration of an all-optical quantum controlled-not gate. *Nature* **426**, 264–267 (2003).

21. Pittman, T. B., Fitch, M. J., Jacobs, B. C. & Franson, J. D. Experimental controlled-not logic gate of single photons in the coincidence basis. *Phys. Rev. A* **68**, 032316 (2003).
22. Gasparoni, S., Pan, J.-W., Walther, P., Rudolph, T. & Zeilinger, A. Realization of a photonic controlled-NOT gate sufficient for quantum computation. *Phys. Rev. Lett.* **92**, 020504 (2004).
23. Sanaka, K., Jennewein, T., Pan, J.-W., Resch, K. & Zeilinger, A. Experimental nonlinear sign-shift for linear optics quantum computation. *Phys. Rev. Lett.* **92**, 017902 (2004).
24. Nielsen, M. A. Optical quantum computation using cluster states. *Phys. Rev. Lett.* **93**, 040503 (2004).
25. Brown, D. E. & Rudolph, T. Efficient linear optical quantum computation. Preprint at (<http://arXiv.org/quant-ph/0405157>) (2004).
26. Walther, P. *et al.* De Broglie wavelength of a non-local four-photon state. *Nature* **429**, 158–161 (2004).
27. Hein, M., Eisert, J. & Briegel, H.-J. Multi-party entanglement in graph states. *Phys. Rev. A* **69**, 062311 (2004).
28. Roos, C. F. *et al.* Control and measurement of three-qubit entangled states. *Science* **304**, 1478–1480 (2004).
29. Weinstein, Y. *et al.* Quantum process tomography of the quantum Fourier transform. *J. Chem. Phys.* **121**, 6117–6133 (2004).
30. Hradil, Z. Quantum-state estimation. *Phys. Rev. A* **55**, R1561–R1564 (1997).
31. Banaszek, K., Ariano, A., Paris, M. & Sacchi, M. Maximum-likelihood estimation of the density matrix. *Phys. Rev. A* **61**, 010304 (1999).
32. James, D., Kwiat, P., Munro, W. & White, A. Measurement of qubits. *Phys. Rev. A* **64**, 052312 (2001).
33. Toth, G. & Guehne, O. Detecting genuine multipartite entanglement with two local measurements. Preprint at (<http://arXiv.org/quant-ph/0405165>) (2004).
34. Dür, W. & Briegel, H.-J. Stability of macroscopic entanglement under decoherence. *Phys. Rev. Lett.* **92**, 180403 (2004).
35. Greenberger, D. M., Horne, M. A. & Zeilinger, A. in *Bell's Theorem, Quantum Theory and Concepts of the Universe* (ed. Kafatos, M.) (Kluwer, Dordrecht, 1989).
36. Zeilinger, A., Horne, M. & Greenberger, D. in *Squeezed States and Quantum Uncertainty* (eds Han, D., Kim, Y. S. & Zachary, W. W.) (NASA Conference Publication 3135, NASA, College Park, 1992).
37. Dür, W., Vidal, G. & Cirac, J. I. Three qubits can be entangled in two inequivalent ways. *Phys. Rev. A* **62**, 62314–62325 (2000).
38. SenDe, A., Sen, U., Wiesniak, M., Kaszlikowski, D. & Zukowski, M. Multi-qubit W states lead to stronger nonclassicality than Greenberger-Horne-Zeilinger states. *Phys. Rev. A* **68**, 623306 (2003).
39. Horodecki, M., Horodecki, P. & Horodecki, R. Separability of mixed states: necessary and sufficient conditions. *Phys. Lett. A* **223**, 1–8 (1996).
40. Peres, A. Separability criterion for density matrices. *Phys. Rev. Lett.* **77**, 1413–1415 (1996).
41. Chuang, I. L. & Nielsen, M. A. Prescription for experimental determination of the dynamics of a quantum black box. *J. Mod. Opt.* **44**, 2455–2467 (1997).
42. Poyatos, J. F., Cirac, J. I. & Zoller, P. Complete characterization of a quantum process: the two-bit quantum gate. *Phys. Rev. Lett.* **78**, 390–393 (1997).
43. Schumacher, B. Quantum coding. *Phys. Rev. A* **51**, 2738–2747 (1995).
44. Coffman, V., Kundu, J. & Wootters, W. K. Distributed entanglement. *Phys. Rev. A* **61**, 052306 (2000).
45. Horodecki, R., Horodecki, P. & Horodecki, M. Violating Bell inequality by mixed spin-1/2 states: necessary and sufficient condition. *Phys. Lett. A* **200**, 340–344 (1995).
46. Ahn, J., Weinacht, T. C. & Bucksbaum, P. H. Information storage and retrieval through quantum phase. *Science* **287**, 463–465 (2000).
47. Bhattacharya, N., van Linden van den Heuvel, H. B. & Spreeuw, R. J. C. Implementation of quantum search algorithm using classical Fourier optics. *Phys. Rev. Lett.* **88**, 137901 (2002).
48. Chuang, I. L., Gershenfeld, N. & Kubinec, M. Experimental implementation of a fast quantum searching. *Phys. Rev. Lett.* **80**, 3408–3411 (1997).
49. Jones, J. A., Mosca, M. & Hansen, R. H. Implementation of a quantum search algorithm on a quantum computer. *Nature* **393**, 344–346 (1998).
50. Kwiat, P. G. *et al.* New high-intensity source of polarization-entangled photon pairs. *Phys. Rev. Lett.* **75**, 4337–4342 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. J. Briegel, D. Browne and M. Zukowski for theoretical discussions, and C. Fürst for assistance with graphics. This work was supported by the Austrian Science Foundation (FWF), NSERC, the European Commission under project RAMBOQ, and by the Alexander von Humboldt Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.W. (pwalth@quantum.at) or A.Z. (zeilinger-office@quantum.at).

Crystal structure of a membrane-bound metalloenzyme that catalyses the biological oxidation of methane

Raquel L. Lieberman & Amy C. Rosenzweig

Departments of Biochemistry, Molecular Biology, and Cell Biology and of Chemistry, Northwestern University, Evanston, Illinois 60208, USA

Particulate methane monooxygenase (pMMO) is an integral membrane metalloenzyme that catalyses the conversion of methane to methanol. Knowledge of how pMMO performs this extremely challenging chemistry may have an impact on the use of methane as an alternative energy source by facilitating the development of new synthetic catalysts. We have determined the structure of pMMO from the methanotroph *Methylococcus capsulatus* (Bath) to a resolution of 2.8 Å. The enzyme is a trimer with an $\alpha_3\beta_3\gamma_3$ polypeptide arrangement. Two metal centres, modelled as mononuclear copper and dinuclear copper, are located in soluble regions of each pmoB subunit, which resembles cytochrome *c* oxidase subunit II. A third metal centre, occupied by zinc in the crystal, is located within the membrane. The structure provides new insight into the molecular details of biological methane oxidation.

The direct conversion of natural gas, composed primarily of methane, to useful fuels and chemicals represents a major challenge for chemistry and engineering. Efficient transformation of methane to a liquid such as methanol would make the extensive and under-used natural gas reserves a realistic alternative to petroleum, affecting worldwide energy use, the environment, and the economy¹. This chemical reaction is difficult to perform because methane is the most inert hydrocarbon (104 kcal mol⁻¹ C–H bond) and the desired product, methanol, reacts further at high temperature². Catalysts that overcome these obstacles have not yet been developed. By contrast, in nature, methane monooxygenase (MMO) enzymes convert methane to methanol at ambient temperature.

MMO is the first enzyme in the metabolic pathway of methanotrophs, which are bacteria that use methane as their sole source of carbon and energy³. Methanotrophs have been used to combat emissions of methane⁴, a potent greenhouse gas, and are also useful for bioremediation because they can oxidise halogenated hydrocarbons^{3,5}. Membrane-bound particulate MMO (pMMO), soluble MMO (sMMO)⁶, and the related enzyme ammonia monooxygenase (AMO)⁷ are the only known enzymes capable of methane hydroxylation. All methanotrophs produce pMMO, which is housed in intracytoplasmic membranes. Under copper-limiting conditions, several strains also produce sMMO³. The well-studied sMMO system comprises a hydroxylase (MMOH), a reductase and a regulatory protein⁶. The crystal structure of MMOH, which contains a carboxylate-bridged di-iron centre, has been known for a decade⁸. By contrast, most questions surrounding the biochemistry, structure and mechanism of the predominant methane oxidation enzyme, pMMO, have remained unanswered despite considerable research efforts in the last 20 years (ref. 9).

pMMO is composed of three subunits, pmoB (α , ~47 kDa), pmoA (β , ~24 kDa)¹⁰, and pmoC (γ , ~22 kDa)¹¹, each containing predicted membrane-spanning helices. The molecular mass and oligomerization state of pMMO are not established, but ~100-kDa $\alpha\beta\gamma$ (ref. 12) and 200-kDa $\alpha_2\beta_2\gamma_2$ (ref. 13) polypeptide arrangements have been proposed. The metal content of pMMO is controversial, with reported values of 2–15 copper ions^{12–16} and 0–2 iron ions^{13–16} per ~100-kDa purified pMMO.

Characterization of pMMO has led to three different models of the metal centre(s). Five trinuclear copper centres have been

proposed by Chan *et al.* on the basis of the hyperfine splitting pattern of an electron paramagnetic resonance (EPR) signal at $g = 2.06$, and have been hypothesized to fall into catalytic (C) and electron transfer (E) functional classes¹⁷. Other researchers, including ourselves, observe a type 2 mononuclear Cu(II) EPR signal only^{13,14,16,18}. DiSpirito and co-workers have suggested two copper and two iron ions with 6–8 additional copper ions bound to methanobactin, a siderophore-like molecule¹⁹, based on pMMO metal content and the isolation of methanobactin¹⁴. In addition to a mononuclear copper site, we have proposed a copper-containing cluster on the basis of extended X-ray absorption fine structure (EXAFS) data best-fitted with a 2.57-Å Cu...Cu interaction^{9,13}. To address these conflicting models and to provide a foundation for directed functional studies, we crystallized *Methylococcus capsulatus* (Bath) pMMO and report here the 2.8-Å resolution structure.

Overall architecture

The pMMO structure was solved by copper single-wavelength anomalous dispersion (SAD) (Supplementary Table S1 and Methods) and refined to 2.8-Å resolution. Three copies each of the pmoA, pmoB and pmoC subunits form a cylindrical $\alpha_3\beta_3\gamma_3$ trimer ~105 Å long and ~90 Å in diameter (Fig. 1a, b). A soluble region composed mainly of six β -barrel structures, two from each protomer, extends ~45 Å away from the membrane, and is supported by 42 transmembrane (TM) helices, 14 from each protomer. A hole is formed in the centre of the trimer (Fig. 1b). In the soluble region, this opening is ~11 Å in diameter and is lined with glutamic acid, aspartic acid and arginine residues. Although these residues stabilize the trimer, they are not conserved among pMMO homologues. The opening then extends into the membrane where it is lined with hydrophobic residues and widens to ~22 Å at the end opposite the soluble region.

The trimeric structure of pMMO was not anticipated and provides the first experimental evidence for a 1:1:1 subunit ratio. The molecular mass of multisubunit membrane proteins is difficult to determine accurately owing to the presence of detergents²⁰. Using nondenaturing gel electrophoresis techniques, we previously suggested a molecular mass of ~200 kDa for purified pMMO, probably corresponding to an $\alpha_2\beta_2\gamma_2$ polypeptide arrangement.

We also observed an ~400-kDa species under some conditions¹³. Chan and co-workers obtained a mass of ~220 kDa by analytical gel filtration, which they attributed to an ~100-kDa $\alpha\beta\gamma$ monomer in complex with 240 detergent molecules¹². Early electron micrographs of reconstituted membranes and two-dimensional crystals of pMMO from *M. capsulatus* (strain M) show a hexagonal particle ~9 nm in diameter with six maxima of protein density²¹. These maxima probably correspond to the six β -barrel structures in the soluble region of *M. capsulatus* (Bath) pMMO (Fig. 1b), suggesting that the crystallographically observed trimer is physiologically relevant.

Protomer structure

Each protomer in the trimer comprises single copies of the pmoB, pmoA and pmoC subunits (Fig. 2a). Consistent with N-terminal sequencing results^{10,13}, pmoB includes residues 33–414. The first 32 residues are proposed to be a leader sequence¹⁰. The soluble regions are derived primarily from pmoB and include two antiparallel β -barrel structures, one at the N terminus and the other at the C terminus (Fig. 2b). The N-terminal β -barrel is composed of seven strands, and is oriented ~90° from the eight-stranded C-terminal β -barrel. The two β -barrel structures are separated by a β -hairpin followed by two TM helices. A 22-residue loop links the TM helices to the C-terminal β -barrel. Residues from this loop participate in trimer interface interactions with the β -barrel structures from the adjacent protomer, whereas the β -hairpin is involved in intraprotomer stabilization. Notably, subunit II of cytochrome *c* oxidase, which contains the dinuclear Cu_A site, also comprises a soluble C-terminal β -barrel domain and two TM helices^{22,23} (Fig. 3). The pmoB subunit houses a dinuclear copper centre as well, in the N-terminal β -barrel (see below). The pmoB and subunit II β -barrels are not superimposable, however.

The pmoA and pmoC subunits reside primarily in the membrane. The pmoA subunit consists of seven TM helices and packs against the two TM helices from pmoB (Fig. 2a, c). These helices, which span the range of observed TM helix lengths²⁴, are inclined with respect to one another, and several are quite tilted with respect to the lipid bilayer normal. The two C-terminal helices face the opening at the trimer centre and interact with their counterparts from the other two protomers. A short helix and a β -hairpin structure protrude from the membrane, and interact with the soluble region of pmoB. pmoC comprises five TM helices that are oriented approximately parallel to the membrane normal and to one another (Fig. 2d). The four N-terminal TM helices average ~29 residues.

Metal centres

The crystal structure reveals three metal centres per protomer (Figs 2a and 4). The first site (Fig. 4a) is located in pmoB ~25 Å above the membrane and near the surface of the N-terminal β -barrel. The metal ion was assigned to be copper from the strong peaks in anomalous Fourier maps calculated using data collected near the copper ('CuANOM' in Supplementary Table S1) and zinc ('Highres' in Supplementary Table S1) absorption edges. Because the zinc absorption edge is higher in energy than the copper absorption edge, anomalous peaks attributable to copper appear in both maps (Fig. 4a). There is no evidence to suggest that this site contains more than one copper ion, and attempts to model additional copper ions were not consistent with anomalous Fourier maps. The copper ion is coordinated by the δ nitrogen atoms of His 48 and His 72 with a nearly linear geometry. Gln 404 is also within 3 Å of the copper ion. Whereas His 72 is highly conserved, His 48 is replaced with asparagine in *M. trichosporium* OB3b and *Methylocystis* sp. strain M pMMO. The assignment of this residue to be histidine is supported by the electron density and the presence of histidine in both the original pmoB gene sequence¹⁰ and the pmoB gene sequence from the *M. capsulatus* (Bath) genome²⁵. The codons for histidine and asparagine differ by a single nucleotide, so the difference may be due to a sequencing error. Alternatively, asparagine could coordinate the copper ion in the homologues.

This mononuclear copper centre resembles the type 2 site found in multicopper oxidases such as ascorbate oxidase and ceruloplasmin²⁶. In these enzymes, the copper ion is coordinated by the ϵ nitrogen atoms of two histidines and a water molecule, visible in all the structures except for the 3.1-Å resolution structure of ceruloplasmin. Solvent ligands may also be present in the pMMO site, but are not detected at 2.8-Å resolution. In accordance with these structural similarities, this copper centre probably gives rise to the type 2 Cu(II) EPR signal observed for purified pMMO^{13,14,16}. In the multicopper oxidases, the type 2 site is located proximal to a dinuclear type 3 site, forming a trinuclear cluster in which the copper ions are within 4–5 Å of one another²⁶. The type 3 site is also ligated by histidines. The surroundings of the pMMO mononuclear site indicate that this site is not a crystallographically depleted variant of the trinuclear multicopper oxidase site. There are no other histidine residues nearby, and the only potential metal binding residues, Gln 404 and Glu 75, are located near the protein surface ~2.7 Å and ~4 Å from the copper centre, respectively. These two residues, which are not conserved, are hydrogen-bonded to one another, and Glu 75 further interacts with pmoB in the neighbouring protomer.

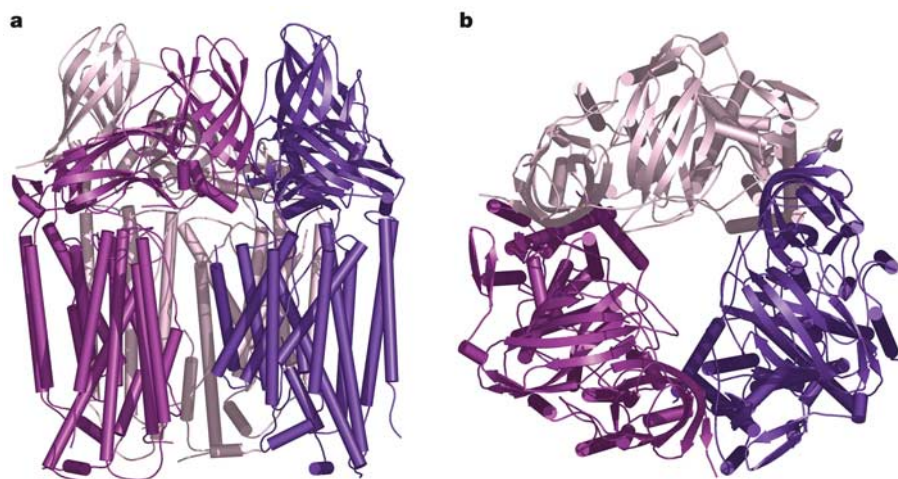


Figure 1 The pMMO trimer viewed parallel to the membrane normal (a) and perpendicular to the membrane normal (b). The three protomers are shown in dark purple, magenta, and light pink. Helices are represented as cylinders and β strands are represented as arrows.

The second copper site (Fig. 4b) is also contained within the N-terminal β -barrel of pmoB, ~ 21 Å from the mononuclear site (Figs 2a and 4). The location of this site, ~ 10 Å above the lipid bilayer interface, is similar to that of the cytochrome *c* oxidase Cu_A site^{22,23} (Fig. 3). We have modelled this site in pMMO as dinuclear on the basis of several observations. First, the electron density in the 3.0-Å resolution copper anomalous Fourier map has an oblong shape. Second and most important, two distinct peaks are observed at this site in a 2.8-Å resolution zinc anomalous Fourier map (Fig. 4b). The same map calculated to 3.0-Å resolution exhibits a single peak, suggesting that the resolution limits of the data are hindering detection of more than one copper ion.

Consistent with the 2.57-Å Cu–metal interaction determined by EXAFS¹³, the Cu–Cu distance in the dinuclear model refines to ~ 2.6 Å. If this site does indeed correspond to the site giving rise to the 2.57-Å Cu–metal interaction, the crystal structure confirms that the interacting metal ion is copper, not iron. No electron density was detected in iron anomalous Fourier maps generated with data sets collected at the iron absorption edge. Although we believe the dinuclear assignment is the best interpretation of the current data, it is also possible to refine the site with a single copper ion. By contrast, modelling three copper ions resulted in significant negative density in $F_o - F_c$ maps, indicating that this site does not

house the trinuclear cluster proposed by Chan and co-workers¹⁷. The oxidation state of the copper ions cannot be discerned from the structure, but a fluorescence scan of the crystal at the copper absorption edge exhibits a peak at $\sim 8,984$ eV. This characteristic Cu(I) feature was observed previously in our X-ray absorption near-edge structure (XANES) spectra¹³, and suggests that the crystals contain some Cu(I). If the mononuclear site generates the type 2 Cu(II) EPR signal, one or both of the copper ions in the dinuclear site is probably Cu(I).

In the dinuclear site, one copper ion is coordinated by the N-terminal residue of pmoB, His 33. Both the N-terminal amino nitrogen and the sidechain δ nitrogen are within coordinating distance. The second copper ion is ligated by the δ nitrogen of His 137 and the ϵ nitrogen of His 139 (Fig. 4b). These residues are highly conserved in pMMO and AMO from a number of organisms. Residues His 33 and His 139 are held in position by hydrogen bonds to the side chain of Glu 35 and the carbonyl oxygen of Gly 152, respectively. Both of these residues are also conserved. Additional terminal or bridging ligands may be present, but are not observed in the 2.8-Å-resolution electron density maps.

Coordination via the N-terminal amine is unusual but not unprecedented. In the bacterial nickel superoxide dismutase, bidentate chelation by the N-terminal amine and histidine side chain is

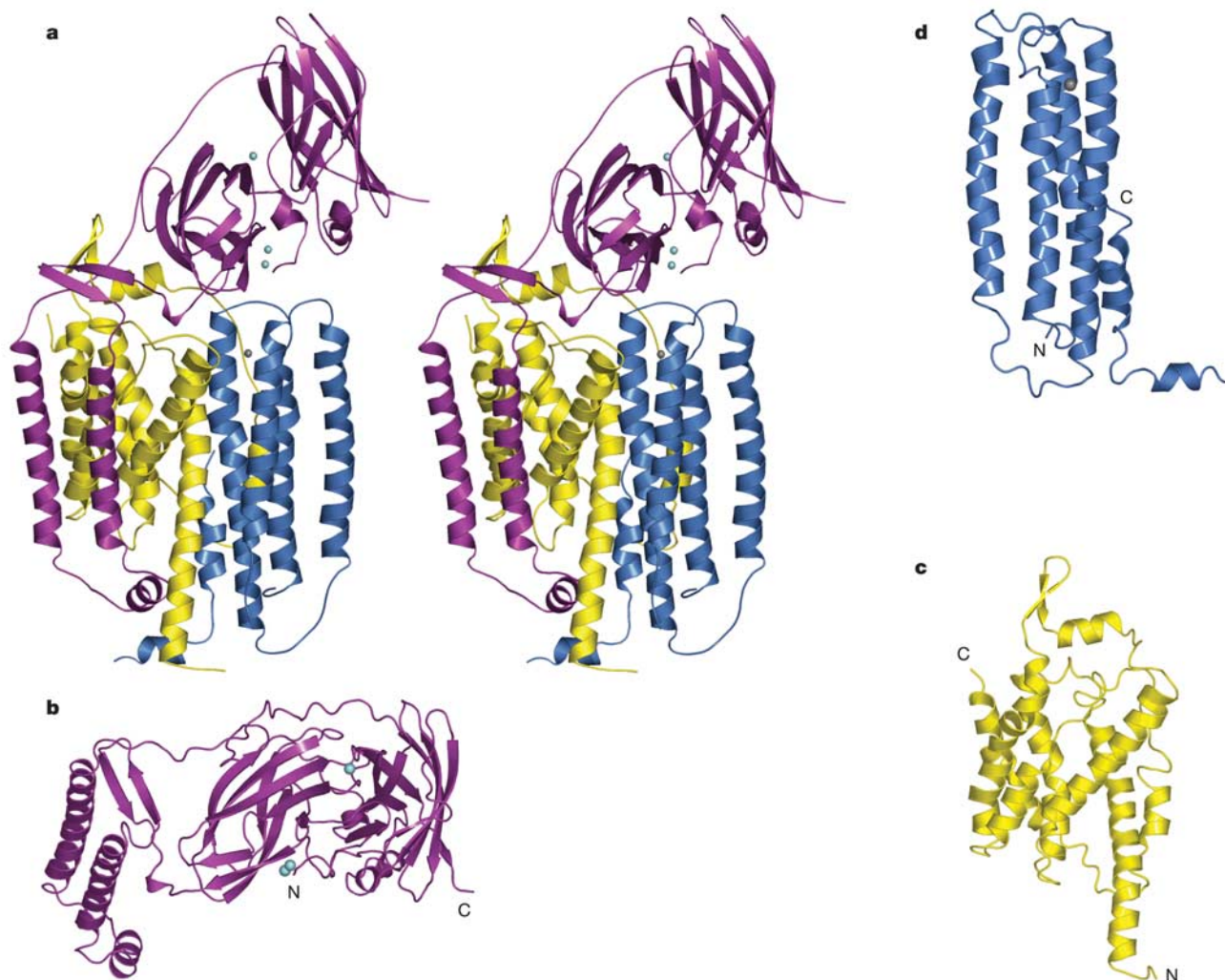


Figure 2 The pMMO subunits. **a**, Stereoview of a single protomer with pmoB shown in magenta, pmoA shown in yellow, and pmoC shown in blue. Copper ions are shown as cyan spheres, and a zinc ion is shown as a grey sphere. **b**, The pmoB subunit viewed

looking down the membrane normal. The N-terminal β -barrel is in the middle. **c**, The pmoA subunit. **d**, The pmoC subunit.



Figure 3 Comparison of pmoB (magenta) with *Paracoccus denitrificans* cytochrome *c* oxidase subunit II (green, PDB accession code 1AR1). The view was generated by superimposing the transmembrane helices and then translating the two proteins away from one another. The N-terminal pMMO β -barrel and the cytochrome *c* oxidase β -barrel are oriented differently. The C-terminal β -barrel of pmoB is transparent for clarity. Copper ions are shown as cyan spheres.

observed. In the proposed reaction mechanism, axial binding of the histidine δ nitrogen to Ni(III) tunes the reduction potential for superoxide dismutation²⁷. Activation of the enzyme requires proteolytic cleavage of a 14-residue leader sequence upstream of this histidine. Similarly, the N-terminal amino nitrogen of the photosynthetic cytochrome *f* subunit is axially bound to a haem iron²⁸. Analogous to cytochrome *f*, assembly of the pMMO dicopper site most probably occurs after the signal peptide is cleaved, when pmoB has already been translocated to the lipid bilayer. Finally, amino-terminal coordination of Cu(II) occurs in the amino-terminal Cu(II)- and Ni(II)-binding (ATCUN) motif, observed in albumin²⁹. We note, however, that the pmoB N terminus is not involved in copper binding if the site is modelled as mononuclear.

The third metal centre (Fig. 4c), modelled as a single zinc ion, is located within the lipid bilayer, ~ 19 Å from the dinuclear copper centre (Figs 2a and 4). The three related sites in the pMMO trimer represent the highest peaks in a zinc anomalous Fourier map, and are not present in anomalous maps calculated using data collected at the lower-energy copper edge. Several additional zinc ions mediate crystal lattice contacts. The zinc ion is coordinated in a distorted tetrahedral geometry by conserved residues Asp 156, His 160 and His 173 from pmoC, and Glu 195 from pmoA (Fig. 4c). This glutamic acid residue can only be tentatively assigned owing to a disordered loop and poor side-chain electron density for residues 193–244. Alternate ligands include Glu 199 and Met 201 from pmoA. Metal analysis by inductively coupled plasma atomic emission spectroscopy (ICP-AES) indicates that pMMO samples contain less than 0.2 moles of zinc per 100 kDa before crystallization. The zinc in this site is therefore derived from zinc acetate in the crystallization buffer. Although this site could be adventitious, it is probably occupied by another metal ion *in vivo* such as copper or iron.

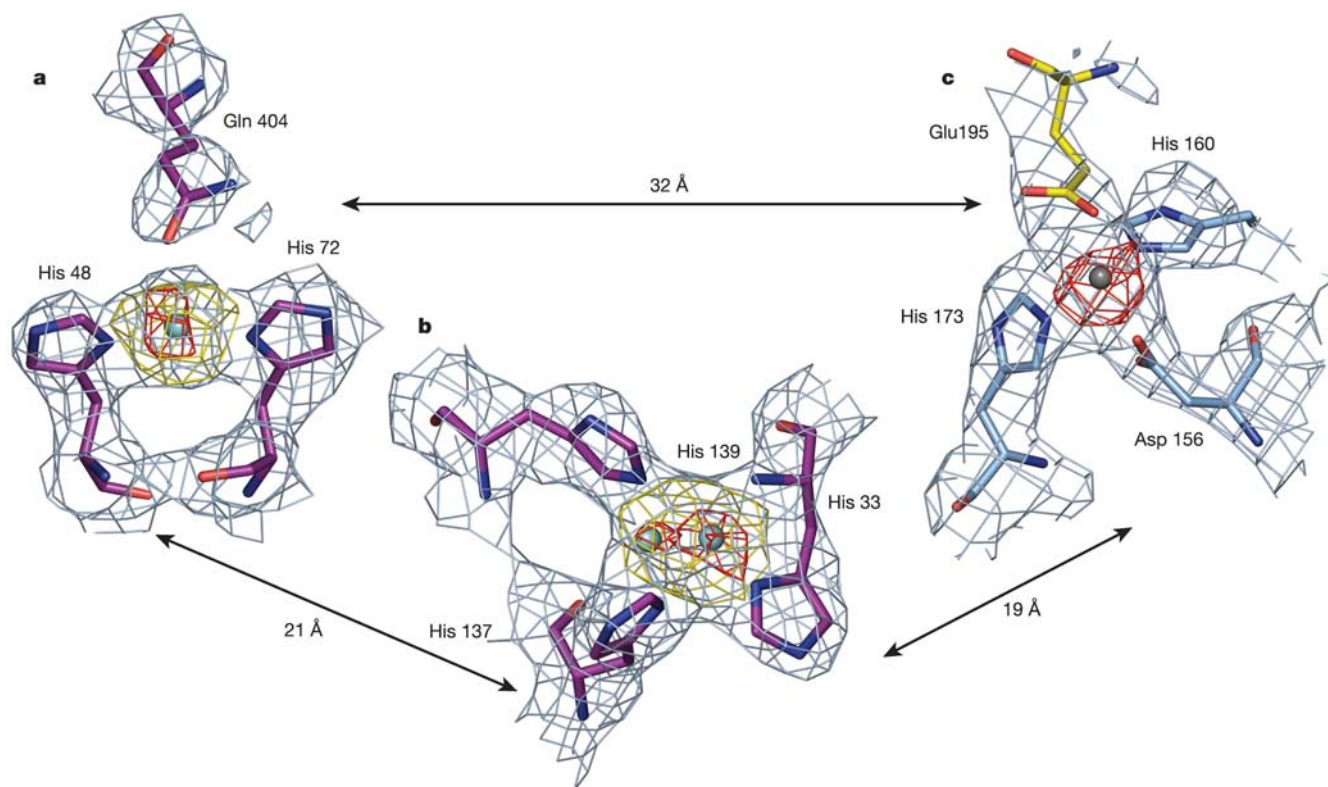


Figure 4 The pMMO metal centres viewed approximately 90° from the orientation shown in Fig. 2a. The distances are measured between metal ions. Anomalous difference Fourier maps calculated using data collected near the copper absorption edge (yellow,

'CuANOM', contoured at 4σ) and near the zinc absorption edge (red, 'Highres', contoured at 4σ) are superimposed on the final $2F_o - F_c$ electron density map (light blue, contoured at 1σ). **a**, The mononuclear copper site. **b**, The dinuclear copper site. **c**, The zinc site.

Taken together, the three metal centres give a stoichiometry of three copper ions and a fourth yet-to-be-identified metal ion per 100 kDa $\alpha\beta\gamma$ protomer, consistent with values measured previously for *M. capsulatus* (Bath) pMMO by us¹³ and Dalton and co-workers¹⁶, but significantly different from the 12–15 copper ions reported by Chan and co-workers¹⁷. One potential explanation is that additional copper binding sites, not occupied in the crystal, are present. Histidine ligation of the trinuclear C-clusters was predicted based on electron spin echo modulation spectroscopic (ESEEM) data for *M. capsulatus* (Bath) membrane samples³⁰. An inspection of all the histidine residues, both conserved and nonconserved, in the pMMO protomer does not reveal any obvious metal binding sites. There is one cluster of primarily hydrophilic residues ~ 13 Å from the zinc site. This region includes conserved residues His 38, Met 42, Asp 47, Asp 49 and Glu 100 from pmoA and Glu 154 from pmoC. These residues may be involved in protomer stabilization, but it is conceivable that they could form a quinone or metal binding site. It is most likely, however, that the observed metal centres represent the only metal binding sites in pMMO.

Functional implications

Oxygen binding, activation and subsequent substrate oxidation have been studied extensively for copper proteins and synthetic model complexes³¹. The nuclearity of biological copper centres is not correlated with reactivity, leaving open the possibility that any one of the pMMO sites could be catalytic. The dinuclear centre (Fig. 4b) is an attractive candidate for the active site because it is unusual. Other dicopper centres, including those in haemocyanin, tyrosinase, and catechol oxidase, can bind and/or activate dioxygen. Whereas haemocyanin only binds dioxygen reversibly, the latter two enzymes hydroxylate tyrosine and convert catechol to quinone, respectively. The Cu–Cu separation in these enzymes is significantly greater than that in pMMO, and each copper ion in haemocyanin and catechol oxidase is coordinated by three histidines³¹ rather than one or two in pMMO. In support of methane binding at the dinuclear site, there is an adjacent cavity lined by several conserved hydrophobic residues including Pro 94 from pmoB and Leu 78, Ile 163, and Val 164 from pmoC. A second cavity involving some of the same residues extends almost to the zinc site.

Alternatively, structural similarities between pmoB and cytochrome *c* oxidase subunit II could indicate that the dinuclear site functions in electron transfer. The copper ions in this site are ~ 20 Å from the other metal ions in the same protomer and >40 Å from the metal sites in the other protomers. Within one protomer, the minimum distance between atoms directly coordinated to these two copper ions and atoms directly coordinated to the mononuclear copper and zinc ions is ~ 14 Å, which is reasonable for electron transfer³². Notably, the side chain of conserved residue pmoB Trp 156 shields the dinuclear copper centre from the solvent. An analogous surface tryptophan in cytochrome *c* oxidase²² has been implicated in electron transfer from cytochrome *c* to the Cu_A site³³. The coordination environment of the pmoB dinuclear site differs from that of Cu_A, however. In Cu_A, the two copper ions are bridged by two cysteine sulphur atoms and terminally coordinated by two histidines, a methionine and a carbonyl oxygen. Key features of Cu_A that contribute to the electron transfer function and are not observed in pMMO include complete delocalization of electrons and the high covalency of the Cu–S(Cys) bonds³⁴.

If the dinuclear site is involved in electron transfer, the mononuclear site (Fig. 4a) might be catalytic. The mononuclear copper centres in peptidylglycine α -hydroxylating monooxygenase (PHM) and dopamine β -monooxygenase (D β M) catalyse hydroxylation of a glycine C α –H bond and a dopamine benzylic C–H bond, respectively³⁵. In these enzymes, a Cu(II)-superoxo complex is proposed to abstract a hydrogen atom from the substrate³⁶. The copper ion is ligated by two histidines and a methionine, which is essential for reactivity³⁵. The pMMO site could exhibit similar

reactivity, although a methionine ligand is not present. In PHM, an additional electron is provided by a second mononuclear copper site located ~ 11 Å away.

In addition, we cannot exclude the possibility that the zinc site (Fig. 4c) is the active site. In cytochrome *c* oxidase, the catalytic centre is located ~ 13 Å within the membrane ~ 20 Å from the electron transfer Cu_A site²³. Multiple combinations of the zinc ligands and proximal residues could provide ligands to either a third copper centre or an iron centre. If Met 201 from pmoA is involved, a copper site similar to that in PHM could assemble. There are only two nearby histidines, so a multinuclear copper site seems unlikely at this position. Finally, it has not escaped our attention that a carboxylate-bridged di-iron centre, similar to that in MMOH⁸, could be accommodated at this position. Two additional nearby carboxylate residues, Asp 168 and Glu 176 from pmoC, could complete the coordination sphere. These two residues are conserved among pMMO homologues, but not in the related AMO. Canonical iron binding motifs³⁷ are not present in the pMMO sequence, however, and no spectroscopic data suggestive of a di-iron centre have been obtained.

The physiological source of electrons for any of the pMMO metal sites is unknown. Exogenous quinones have been shown to stimulate NADH-dependent pMMO activity in solubilized membranes. Shiemke and co-workers have proposed that a type 2-NADH:quinone oxidoreductase (NDH-2) uses NADH generated by oxidation of formate and formaldehyde to reduce endogenous quinones. These resultant quinols could then reduce pMMO either directly or through an additional yet-to-be-identified reductase³⁸. There is a nonprotein patch of strong electron density near the interface between the two pmoB β -barrels and ~ 13 Å from both the copper centres that could accommodate a quinol (Supplementary Fig. S1). The structure also reveals a possible docking site for NDH-2 or other reductases, a negatively charged patch on the outer surface of the soluble domains of pmoB close to the dinuclear copper site (Supplementary Fig. S2). This region includes conserved residues pmoB Glu 35, pmoB Asp 368, and pmoC Asp 45. A cluster of acidic residues also forms the proposed docking site for cytochrome *c* on cytochrome *c* oxidase³⁹.

In conclusion, the pMMO structure reveals an unexpected trimeric arrangement and the overall folds of the three subunits. Two of the three metal centres, modelled as mononuclear and dinuclear copper, are located within the soluble regions of the pmoB subunit. The third metal centre, occupied by zinc in the crystal, lies within the membrane with ligands derived from both pmoC and pmoA. Direct electron transfer between metal centres may be possible. Neither the site of methane oxidation nor the pathway(s) of substrate entry and product egress have yet been identified. Studies to address these issues are underway, and will provide a foundation for future studies of the pMMO methane oxidation mechanism. □

Methods

Protein purification and crystallization

Methylococcus capsulatus (Bath) cultures were fermented and the membranes isolated as described previously¹³. pMMO was purified in the presence of the detergent undecyl- β -D-maltoside (Anatrace) by ammonium sulphate precipitation followed by anion exchange chromatography on a Source 30Q (Amersham Biosciences) column. Purified pMMO was exchanged into 50 mM Hepes pH 7.5 containing 0.12% Cymal-5 (Anatrace) by fourfold dilution and concentration using a microcon YM-100 (Millipore). As described previously, most samples were not active¹³. Crystals were grown at room temperature in clover-leaf sitting drop trays (DeCode Biostructures) by mixing 1 μ l of ~ 20 mg ml^{−1} pMMO with 1 μ l precipitant solution containing 200 mM zinc acetate and 8–12% PEG 8000. Colourless rod-shaped crystals with final dimensions of $\sim 100 \times 100 \times 400$ μ m appeared after 1–2 weeks. Activity assays were not attempted on the crystals owing to insufficient quantities. Crystals were flash-frozen using 15% ethylene glycol as a cryoprotectant.

Structure determination and refinement

The structure was solved by copper SAD (Supplementary Table S1). All data sets were collected at the DND-CAT beamline at the Advanced Photon Source using a 3K $\times$ 3K Mar

charge-coupled device (CCD) detector and processed with XDS⁴⁰ and SCALA⁴¹. Six copper sites were located using SOLVE⁴². A figure of merit of 0.24 was obtained for data in the 30–4.0-Å resolution range. High redundancy was critical to locating the copper sites, and was achieved by merging three complete data sets collected at the same energy on the same crystal ('CuSAD' in Supplementary Table S1). An interpretable electron density map was obtained from RESOLVE⁴³ following threefold noncrystallographic symmetry averaging, density modification using a solvent content of 70%, and phase extension to 2.8 Å resolution ('Highres' in Supplementary Table S1). Model building and adjustment was performed with XtalView⁴⁴. Approximately 85 iterative refinement cycles with CNS⁴⁵ followed by model rebuilding were performed. Translation liberation screw-rotation (TLS) parameters and restrained refinement options in REFMAC5⁴⁶ were used for the final refinement cycles. Noncrystallographic symmetry restraints were imposed throughout the refinement. The asymmetric unit contains one pMMO trimer.

The model includes all three pMMO subunits, pmoA, pmoB and pmoC, numbered according to the translated protein sequences for the genes *pmoA2*, *pmoB2* and *pmoC2*¹. Few sequence differences exist between the two pMMO gene copies in *M. capsulatus* (Bath). We chose the second gene copy based on the electron density for residue 173 in a well-defined region of pmoA. The density at this position is more consistent with tyrosine (*pmoA2*) than asparagine (*pmoA*). Several regions of pmoA and pmoC proved difficult to model, including the six N-terminal residues of pmoA and the 44 N-terminal residues of pmoC, which may be a leader sequence. Owing to a disordered loop comprising residues 193–195 and poor side chain density, residues 193–244 of pmoA are tentatively assigned, and the final 3 C-terminal residues are not modelled. In pmoC, residues 204–230 and 260–289 are not visible, and residues 231–259 are tentatively assigned. The final model consists of 2,424 amino acid residues, nine copper ions, and eight zinc ions. A Ramachandran plot calculated with PROCHECK⁴¹ indicates that 95.6%, 97.2% and 97.1% of the residues in pmoA, pmoB and pmoC, respectively, are in the most favoured and additional allowed regions. Figures were generated with PyMOL⁴⁷ and GRASP⁴⁸.

Received 9 November 2004; accepted 5 January 2005; doi:10.1038/nature03311.

Published online 26 January 2005.

- Periana, R. A. *et al.* Perspectives on some challenges and approaches for developing the next generation of selective, low temperature, oxidation catalysts for alkane hydroxylation based on the CH activation reaction. *J. Mol. Catal. A* **220**, 7–25 (2004).
- Periana, R. A. *et al.* A mercury-catalyzed, high-yield system for the oxidation of methane to methanol. *Science* **259**, 340–343 (1993).
- Hanson, R. S. & Hanson, T. E. Methanotrophic bacteria. *Microbiol. Rev.* **60**, 439–471 (1996).
- Park, S., Brown, K. W. & Thomas, J. C. The effect of various environmental and design parameters on methane oxidation in a model biofilter. *Waste Manag. Res.* **20**, 434–444 (2002).
- Sullivan, J. P., Dickinson, D. & Chase, C. A. Methanotrophs, *Methylosinus trichosporium* OB3b, sMMO, and their application to bioremediation. *Crit. Rev. Microbiol.* **24**, 335–373 (1998).
- Merkx, M. *et al.* Dioxygen activation and methane hydroxylation by soluble methane monooxygenase: a tale of two irons and three proteins. *Angew. Chem. Int. Edn Engl.* **40**, 2782–2807 (2001).
- Arp, D. J., Sayavedra-Soto, L. A. & Hommes, N. G. Molecular biology and biochemistry of ammonia oxidation by *Nitrosomonas europaea*. *Arch. Microbiol.* **178**, 250–255 (2002).
- Rosenzweig, A. C., Frederick, C. A., Lippard, S. J. & Nordlund, P. Crystal structure of a bacterial non-haem iron hydroxylase that catalyses the biological oxidation of methane. *Nature* **366**, 537–543 (1993).
- Lieberman, R. L. & Rosenzweig, A. C. Biological methane oxidation: regulation, biochemistry, and active site structure of particulate methane monooxygenase. *Crit. Rev. Biochem. Mol. Biol.* **39**, 147–164 (2004).
- Semrau, J. D. *et al.* Particulate methane monooxygenase genes in methanotrophs. *J. Bacteriol.* **177**, 3071–3079 (1995).
- Stolyar, S., Costello, A. M., Peeples, T. L. & Lidstrom, M. E. Role of multiple gene copies in particulate methane monooxygenase activity in the methane-oxidizing bacterium *Methylococcus capsulatus* Bath. *Microbiology* **145**, 1235–1244 (1999).
- Yu, S. S.-F. *et al.* Production of high-quality particulate methane monooxygenase in high yields from *Methylococcus capsulatus* (Bath) with a hollow-fiber membrane bioreactor. *J. Bacteriol.* **185**, 5915–5924 (2003).
- Lieberman, R. L. *et al.* Purified particulate methane monooxygenase from *Methylococcus capsulatus* (Bath) is a dimer with both mononuclear copper and a copper-containing cluster. *Proc. Natl Acad. Sci. USA* **100**, 3820–3825 (2003).
- Choi, D. W. *et al.* The membrane-associated methane monooxygenase pMMO and pMMO-NADH:quinone oxidoreductase complex from *Methylococcus capsulatus* Bath. *J. Bacteriol.* **185**, 5755–5764 (2003).
- Nguyen, H. H., Elliott, S. J., Yip, J. H. & Chan, S. I. The particulate methane monooxygenase from *Methylococcus capsulatus* (Bath) is a novel copper-containing three-subunit enzyme. Isolation and characterization. *J. Biol. Chem.* **273**, 7957–7966 (1998).
- Basu, P., Katterle, B., Andersson, K. K. & Dalton, H. The membrane-associated form of methane monooxygenase from *Methylococcus capsulatus* (Bath) is a copper/iron protein. *Biochem. J.* **369**, 417–427 (2003).
- Chan, S. I., Chen, K. H.-C., Yu, S. S.-F., Chen, C.-L. & Kuo, S. S.-J. Toward delineating the structure and function of the particulate methane monooxygenase from methanotrophic bacteria. *Biochemistry* **43**, 4421–4430 (2004).
- Lemos, S. S., Yuan, H. & Perille-Collins, M. L. Review of multifrequency EPR of copper in particulate methane monooxygenase. *Curr. Top. Biophys.* **26**, 43–48 (2002).
- Kim, H. J. *et al.* Methanobactin, a copper-acquisition compound from methane oxidizing bacteria. *Science* **305**, 1612–1615 (2004).

- le Maire, M., Champell, P. & Møller, J. V. Interaction of membrane proteins and lipids with solubilizing detergents. *Biochim. Biophys. Acta* **1508**, 86–111 (2000).
- Tsuprun, V. L. *et al.* Electron microscopy of methane monooxygenase of the methane-oxidizing bacterium *Methylococcus capsulatus*. *Dokl. Akad. Nauk SSSR* **292**, 490–493 (1987).
- Iwata, S., Ostermeier, C., Ludwig, B. & Michel, H. Structure at 2.8 Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*. *Nature* **376**, 660–669 (1995).
- Tsukihara, T. *et al.* Structures of the metal sites of oxidized bovine heart cytochrome c oxidase at 2.8 Å. *Science* **269**, 1069–1074 (1995).
- Bowie, J. U. Helix packing in membrane proteins. *J. Mol. Biol.* **272**, 780–789 (1997).
- Ward, N. *et al.* Genomic insights into methanotrophy: the complete genome sequence of *Methylococcus capsulatus* (Bath). *PLoS Biol.* **2**, e303 (2004).
- Solomon, E. I., Sundaram, U. M. & Machonkin, T. E. Multicopper oxidases and oxygenases. *Chem. Rev.* **96**, 2563–2605 (1996).
- Barondeau, D. P., Kassmann, C. J., Bruns, C. K., Tainer, J. A. & Getzoff, E. D. Nickel superoxide dismutase structure and mechanism. *Biochemistry* **43**, 8038–8047 (2004).
- Martinez, S. E., Huang, D. S., Szczepaniak, A., Cramer, W. A. & Smith, J. L. Crystal structure of chloroplast cytochrome *f* reveals novel cytochrome fold and unexpected heme ligation. *Structure* **2**, 95–105 (1994).
- Harford, C. & Sarkar, B. Amino terminal Cu(II)- and Ni(II)-binding (ATCUN) motif of proteins and peptides: metal binding, DNA cleavage, and other properties. *Acc. Chem. Res.* **30**, 123–130 (1997).
- Elliott, S. J., Randall, D. W., Britt, R. D. & Chan, S. I. Pulsed EPR studies of particulate methane monooxygenase from *Methylococcus capsulatus* (Bath): evidence for histidine ligation. *J. Am. Chem. Soc.* **120**, 3247–3248 (1998).
- Mirica, L. M., Ottenwaelder, X. & Stack, T. D. P. Structure and spectroscopy of copper-dioxygen complexes. *Chem. Rev.* **104**, 1013–1045 (2004).
- Gray, H. B. & Winkler, J. R. Electron tunneling through proteins. *Q. Rev. Biophys.* **36**, 341–372 (2003).
- Witt, H., Malatesta, F., Nicoletti, F., Brunori, M. & Ludwig, B. Tryptophan 121 of subunit II is the electron entry site to cytochrome-c oxidase in *Paracoccus denitrificans*. *J. Biol. Chem.* **273**, 5132–5136 (1998).
- George, S. D. *et al.* A quantitative description of the ground-state wave function of Cu_A by X-ray absorption spectroscopy: comparison to plastocyanin and relevance to electron transfer. *J. Am. Chem. Soc.* **123**, 5757–5767 (2001).
- Prigge, S. T., Mains, R. E., Eipper, B. A. & Amzel, L. M. New insights into copper monooxygenases and peptide amidation: structure, mechanism and function. *Cell. Mol. Life Sci.* **57**, 1236–1259 (2000).
- Chen, P. & Solomon, E. I. O₂ activation by binuclear Cu sites: noncoupled versus exchange coupled reaction mechanisms. *Proc. Natl Acad. Sci. USA* **101**, 13105–13110 (2004).
- Nordlund, P. & Eklund, H. Di-iron-carboxylate proteins. *Curr. Opin. Struct. Biol.* **5**, 758–766 (1995).
- Cook, S. A. & Shiemke, A. K. Evidence that a type-2 NADH:quinone oxidoreductase mediates electron transfer to particulate methane monooxygenase in *Methylococcus capsulatus*. *Arch. Biochem. Biophys.* **398**, 32–40 (2002).
- Maneg, O., Malatesta, F., Ludwig, B. & Drosou, V. Interaction of cytochrome c with cytochrome oxidase: two different docking scenarios. *Biochim. Biophys. Acta* **1655**, 274–281 (2004).
- Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.* **26**, 795–800 (1993).
- Collaborative Computational Project Number 4. The CCP4 suite programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
- Terwilliger, T. C. Automated main-chain model building by template matching and iterative fragment extension. *Acta Crystallogr. D* **59**, 34–44 (2002).
- McRee, D. E. XtalView Xfit—a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165 (1999).
- Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular crystallography. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Winn, M. D., Isupov, M. N. & Murshudov, G. N. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D* **57**, 122–133 (2001).
- Delano, W. L. *The PyMOL Molecular Graphics System* (DeLano Scientific, San Carlos, California, 2002).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by a grant from the American Chemical Society Petroleum Research Fund (A.C.R.), funds from the David and Lucile Packard Foundation (A.C.R.), and the NIH (A.C.R.). R.L.L. was supported in part by a NIH training grant. We thank D. Shrestha for assistance cultivating *M. capsulatus* (Bath), M. Sommerhalter for assistance with data collection, J. Brunzelle for suggestions, Z. Wawrzak for assistance with data collection, and B. Hoffman, T. Stemmler, and K. Karlin for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.C.R. (amy@northwestern.edu). Coordinates have been deposited in the Protein Data Bank (accession code 1YEW).

Structural basis of HutP-mediated anti-termination and roles of the Mg^{2+} ion and L-histidine ligand

Thirumananseri Kumarevel¹, Hiroshi Mizuno^{2*} & Penmetcha K. R. Kumar¹

¹Functional Nucleic Acids Group, Institute for Biological Resources and Functions, National Institute of Advanced Industrial Science and Technology (AIST), Central 6, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan

²Department of Biochemistry, National Institute of Agrobiological Sciences, 2-1-2, Kannondai, Tsukuba, Ibaraki 305-8602, Japan

* Present address: NEC Soft. Ltd, 1-18-6, Shinkiba, Koto-ku, Tokyo 106-8608, Japan

HutP regulates the expression of the *hut* structural genes of *Bacillus subtilis* by an anti-termination mechanism and requires two components, Mg^{2+} ions and L-histidine. HutP recognizes three UAG triplet units, separated by four non-conserved nucleotides on the terminator region. Here we report the 1.60-Å resolution crystal structure of the quaternary complex (HutP–L-histidine– Mg^{2+} –21-base single-stranded RNA). In the complex, the RNA adopts a novel triangular fold on the hexameric surface of HutP, without any base-pairing, and binds to the protein mostly by specific protein–base interactions. The structure explains how the HutP and RNA interactions are regulated critically by the L-histidine and Mg^{2+} ion through the structural rearrangement. To gain insights into these structural rearrangements, we solved two additional crystal structures (uncomplexed HutP and HutP–L-histidine– Mg^{2+}) that revealed the intermediate structures of HutP (before forming an active structure) and the importance of the Mg^{2+} ion interactions in the complexes.

Genes responsible for L-histidine degradation and utilization, as carbon and nitrogen sources under nutrient-limiting conditions, are located within the *hut* operon in *Bacillus subtilis*. The *hut* operon consists of five structural genes, *hutH*, *hutU*, *hutI*, *hutG* and *hutM*, and a positive regulatory gene, *hutP*^{1–5}. The *hutP* gene is located just downstream of the promoter, whereas the structural genes are farther downstream³. An overlapping nucleotide sequence, located between the *hutP* gene and the structural genes, is predicted to form an anti-termination/termination structure, depending on the availability of L-histidine^{4,5}. Previous biochemical and genetic studies indicated that the *hut* messenger RNA forms a terminator structure (+498 to +572 nucleotides; Fig. 1a) in the region between the *hutP* and *hutH* genes, and regulates *hut* operon transcription by an anti-termination mechanism. Anti-termination is one of several mechanisms used by various bacteria to regulate the transcription of biosynthetic operons^{6–8}. Several bacterial proteins that regulate their corresponding operons by anti-termination/termination on activation by their respective ligands have been described, including TRAP, PyrR, LacT, BglG, SacT and GlpP^{9–15}. However, HutP has no sequence similarity to the other proteins reported so far, including the RNA-binding proteins. To identify the important chemical groups of L-histidine for HutP activation, we analysed 15 different L-histidine analogues and found that both the imidazole group and the L-histidine backbone moiety are important for the activation^{16,17}. Recently, we solved the crystal structure of HutP with a histidine β -naphthylamide (HBN), which revealed a novel RNA-binding motif¹⁶. Three dimers of HutP are related by three-fold symmetry to form the hexamer, which seems to be necessary for its function. At the dimer interface, a hydrophobic pocket was formed in which the imidazole group of the HBN analogue resided through a hydrogen bond with Glu 81. Residues Glu 81 and Arg 88 also formed a salt bridge, which is required for stabilizing the hydrophobic pocket¹⁶ (Supplementary Fig. S1).

We recently identified a minimal 21-base region for HutP binding in *hut* mRNA, spanning the region between +496 and +515 nucleotides¹⁶. On the basis of *in vitro* selection and site-specific

mutational analyses, we also identified UAGNNNNUAGNNNNUAG as the recognition motif, in which N indicates any base, and the important chemical groups of the bases for HutP recognition within the core region of the UAG motif¹⁸. These analyses suggested that each HutP monomer recognizes one UAG motif, in which, the 2'-OH and 6-amino group of the A base and the 2-amino group of the G base are the most important groups for the protein–RNA interactions, and the spacer nucleotides have no major role in these interactions; at least two UAG motifs are required for their cooperative interactions for protein–RNA recognition¹⁸. However, the spacer lengths have a function in placing the UAG motif at the next binding site in the hexameric HutP.

HutP also exists in three other *Bacillus* species, including *B. anthracis*¹⁹, *B. cereus*²⁰ and *B. halodurans*²¹, with 60% sequence identity. At present, the attenuation/anti-termination protein–RNA complex structures are available for only two proteins, including the full-length TRAP²² and the amino-terminal peptide fragment of LicT²³. However, the structure in the absence of ligand (L-tryptophan) for the full-length TRAP is not available to reveal insights into any structural rearrangement that might occur on binding to their respective ligands. In the present study we sought to determine the anti-terminator complex structure of HutP and to identify the intermediate structures of HutP.

Here we show that HutP requires metal ions for exerting its anti-termination function and reveal the crystal structure of the anti-termination complex. The quaternary complex structure shows how HutP specifically recognizes the conserved sequences within the *hut* mRNA, and reveals the unexpected direct role of the Mg^{2+} ion for mediating the L-histidine-dependent structural rearrangement in the protein. To unravel these structural changes, we have solved two additional crystal structures (uncomplexed HutP and HutP–L-histidine– Mg^{2+}) and found that the Mg^{2+} ion coordinates with the L-histidine to facilitate an appropriate structural rearrangement before its recognition of the cognate RNA. Once HutP has undergone this structural rearrangement, it binds specifically to the RNA sequences within the terminator region and wraps the RNA

around the protein; by this, the RNA structure may reorganize and destabilize the terminator structure.

Requirement of divalent metal ions

Our previous analyses suggested that the HutP protein binds to its cognate RNA only in the presence of L-histidine¹⁶. However, these studies were performed in binding buffer containing 10 mM MgCl₂. When MgCl₂ was omitted from the binding reactions, HutP failed to bind to *hut* mRNA. To analyse the importance of the divalent metal ions in the formation of the HutP–L-histidine–RNA complex, we used a non-denaturing gel-shift assay to monitor the complex. To remove the endogenous metal ions, the HutP protein was dialysed against binding buffer lacking MgCl₂. We used a shorter RNA, 5'-UUUAGUUUUUAGUUUUUAGUU-3', in this analysis than previously^{16,18}. As shown in Fig. 1b, in the absence of MgCl₂,

the HutP protein failed to bind to RNA, even in the presence of L-histidine (lanes 4 and 5). When 0.5 mM MgCl₂ was incorporated in the binding buffer, we clearly observed quaternary complex formation (Fig. 1b, lane 6). The divalent metal ion used in these studies is also present in physiologically relevant concentrations in bacterial cells. Several other divalent cations such as Mn²⁺, Zn²⁺ and Cd²⁺ are also capable of mediating the HutP–RNA interactions (T.S.K., H.M. & P.K.R.K., unpublished data). Thus, HutP requires not only L-histidine but also the divalent metal ions for complex formation.

Structure of the quaternary complex

The overall HutP protein forms the hexamer structure and resembles a flattened cylinder about 48 Å thick along the three-fold axis and 83 Å wide (Fig. 1c–e). The bound 21-mer RNA was

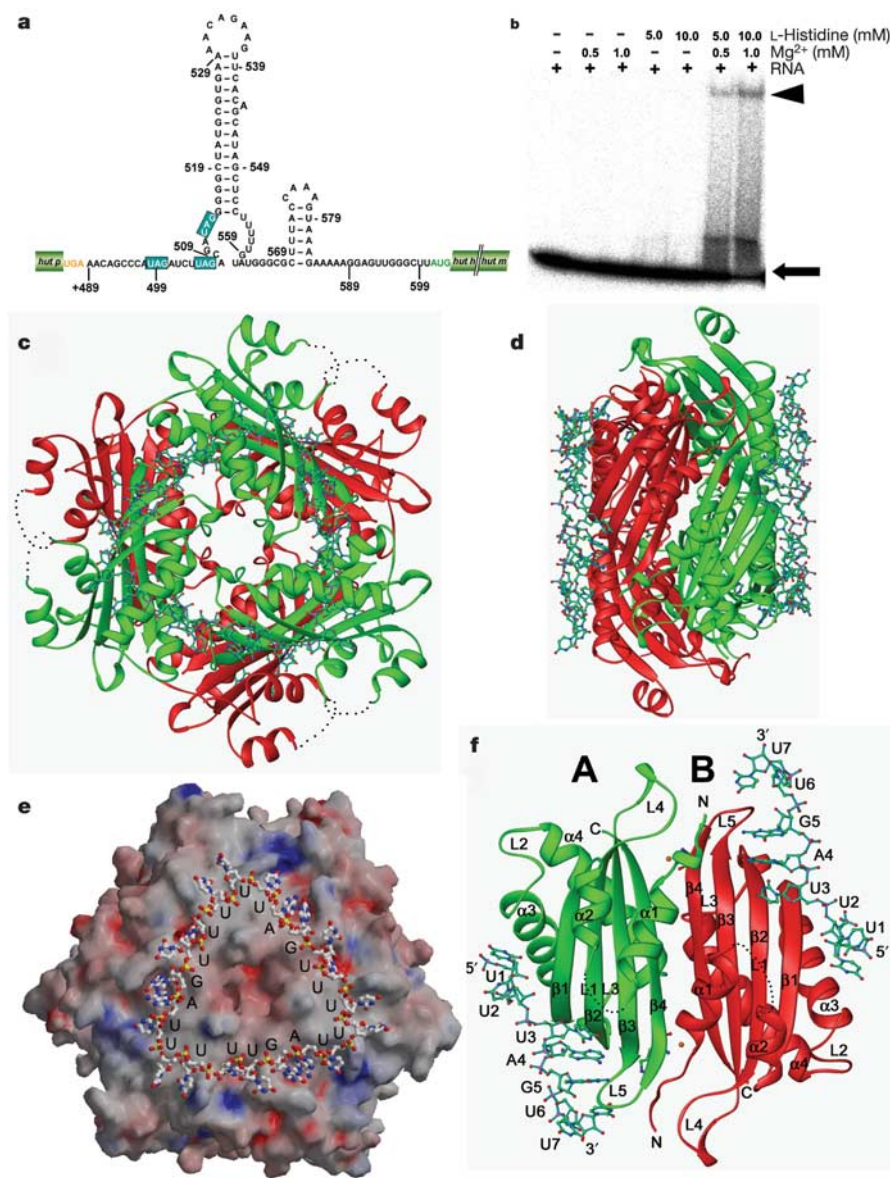


Figure 1 Terminator structure of *hut* mRNA, and the overall structure of the HutP quaternary complex. **a**, The proposed *hut* mRNA terminator structure. In the HutP recognition motif the UAG triplets are shown in a blue box. The stop codon of *hutP* gene is shown in yellow and the start codon of the *hutH* gene is shown in green. **b**, The importance of L-histidine and Mg²⁺ ion for complex formation. Purified HutP (1,200 nM) was mixed with the 21-mer RNA at 20 nM (labelled RNA 10⁴ c.p.m.). Free RNA and complexed RNA are indicated by an arrow and an arrowhead, respectively. **c**, Hexameric HutP quaternary

complex, viewed along the three-fold axis. **d**, Alternative view of the hexamer, rotated by 90° along the y axis. Ball-and-stick models were used for L-histidine and the bound 21-mer RNA. **e**, Molecular surface representation of the quaternary complex and the bound RNA, shown in a ball-and-stick model as in **c**. **f**, HutP dimer viewed along the pseudo-two-fold axis. The bound L-histidine and RNA are shown as a ball-and-stick model and the Mg²⁺ ions are represented by a yellow cpk (Corey–Pauling–Koltun) model.

recognized on both the top and bottom surfaces of the cylinder of the HutP in a previously unreported triangular conformation (Fig. 1e). Thus, the HutP hexamer recognizes two 21-mer RNA molecules; this is consistent with our previous biochemical studies, which indicated that it forms predominantly 1:2 (protein:RNA) complex ratios^{16,18}. This HutP hexamer complex is formed from a three individual dimers related by a crystallographic three-fold axis, with D3 molecular symmetry (Fig. 1c–e). The surface buried by dimer–dimer association within the hexamer HutP has lost an accessible surface area of about 3,700 Å². The crystallographic asymmetric unit of the quaternary complex contains one homodimer of HutP (molecules A and B), two L-histidines, two Mg²⁺ ions and two seven-nucleotide fragments of the 21-mer RNA, which are related by a non-crystallographic two-fold axis (Fig. 1f). In the HutP dimer, L-histidine and Mg²⁺ ion are associated with each other and are located in the dimer interface. The bound L-histidine and Mg²⁺ ions are recognized on one side of the HutP and the RNA is recognized on the other side. Molecules A and B of HutP in the asymmetric unit are essentially identical, with a root-mean-square deviation (r.m.s.d.) of 0.13 Å for superposition of all of the 143 observed Cα atoms, except for the N-terminal residue and four residues in the loop region. Each HutP monomer exhibits the typical structural elements of an α/β protein, arranged in the order α-α-β-α-α-β-β-β in the primary structure, and the four antiparallel β-strands form a β-sheet in the order β1-β2-β3-β4, with two α-helices each on the front (α1 and α2) and at the back (α3 and α4) of the β-sheet (Fig. 1f). The loop regions, which are L1 between α1 and α2, L2 between α3 and α4, L3 between α4 and β2, L4 between β2 and β3, and L5 between β3 and β4, generally protrude from the surface. The surface buried by dimer formation represents an accessible surface area loss of about 1,500 Å² per monomer. Each HutP monomer recognizes one unit of the seven-nucleotide fragment from the 21-mer RNA (5′-UUUAGUU-3′)₃ as it passes across the β-sheet.

HutP–RNA interactions

The bound 21-mer RNA is covalently intact in the crystal, with its three repeated seven-nucleotide sequences arranged symmetrically on each side of the HutP hexamer surface with a triangular shape (Fig. 1c–e). In the seven-nucleotide RNA, the central A4–G5 forms an extensive hydrogen-bonding network with HutP, and a clear electron density along the nucleotide is visible (Fig. 2a, b). U7 interacts with the neighbouring dimer within the HutP hexamer (Fig. 2b). Bases U1, U2 and U6 are partly disordered and show no interactions with HutP. The overall RNA structure adopts an extended A-RNA conformation, with C3′ endo sugar conformations for all of the residues. The central bases A4 and G5 are stacked on each other, and the remaining bases probably serve as spacers without much interaction with the protein molecule; their basic requirement is apparently to place the next UAG binding site on the hexameric HutP. Because the RNA passes across the β-sheet, strands β2 and β3 bend down slightly to form a smooth surface to accommodate the RNA. The β2 and β3 strands interact with the A4 base, β4 contacts the G5 base (Fig. 1f), and the 3′ end of the U7 base interacts with the α3 helix of the neighbouring dimer within the HutP hexamer.

The 2′-OH groups of the A4 and U7 sugars form hydrogen bonds with the side chains of Thr 99 and Thr 56, respectively (Fig. 2b). These two interactions involving the 2′-OH are critical for the protein–RNA interactions, which is consistent with both the deoxyribonucleotide substitutional analysis¹⁸ and the mutational analysis of the HutP protein. This explains the ability of HutP to distinguish between RNA and DNA.

All of the hydrogen-bonding interactions in this protein–RNA complex are mainly directed towards the bases and the sugars, not the phosphates of the backbone, thus readily explaining the sequence-dependent nature of the recognition (Fig. 2b). We showed

previously that Glu 55 and Glu 137 might be responsible for the protein–RNA interactions, and those analyses correlated well with the present quaternary complex structure^{16,18}. To substantiate the residues at the RNA-binding site on HutP, the critical residues interacting with the RNA—Glu 55, Thr 56, Thr 99, Thr 128 and Glu 137—were replaced with Ala, and the protein–RNA interactions were evaluated with both the gel mobility-shift (Fig. 2c) and filter-binding (Supplementary Fig. S2) assays. The Thr99Ala and Glu137Ala mutant proteins failed to bind to the RNA, the Glu55Ala and Thr56Ala mutants showed comparable affinities for the wild-type HutP protein, and the binding ability for the Thr128Ala mutant decreased 15-fold. These studies showed that residues Thr 99 and Glu 137 were essential for RNA binding. These analyses further suggested that the 2′-OH of the A4 ribose interaction with Thr 99 was more important in HutP interactions than the interaction of the 6-amino group of the A4 base with Thr 128 and that this was the reason why A4 could be replaced with U with 30-fold less affinity¹⁸, owing to the loss of the 6-amino interactions with Thr 128. These three residues are completely responsible for the recognition of specific bases on RNA. The other two residues, Glu 55 and Thr 56, have additional contacts with the spacer bases but do not affect the overall binding or the formation of the complex; this is consistent with our analyses^{16,18}.

It is interesting to compare the RNA conformations within both HutP and TRAP²², because these two proteins recognize UAG motifs in their respective mRNAs. The overall conformations of these two RNAs are very different from each other (Supplementary Discussion 1; Supplementary Fig. S3). The distinct feature shared by these two protein–RNA complexes is the absence of direct interactions of the proteins with the RNA phosphates. The most significant difference between the currently known anti-terminator proteins and HutP is in the requirement of metal ions for HutP.

L-Histidine and Mg²⁺-ion-binding site

As shown in Fig. 3a, the L-histidine imidazole ring projects downwards into the solvent and the imidazole nitrogen hydrogen bonds to a water molecule, whereas the amide nitrogen hydrogen bonds to the side-chain oxygen of Tyr 69 of the neighbouring dimer within the HutP hexamer. In contrast, the amino and carboxyl groups of the L-histidine backbone have many interactions as shown. The carboxyl group forms a typical salt bridge with the guanidyl group of Arg 88. In the HutP–HBN complex, this guanidyl group forms a typical salt bridge with the carboxyl group of Glu 81, which is important for the formation of the hydrophobic pocket¹⁶. The amino and carboxyl groups of L-histidine coordinate with the Mg²⁺ ion. The Mg²⁺ ion is involved in the formation of a typical six-coordination sphere. Of the six coordinations, three were with the imidazole nitrogens of His 138, His 73 and His 77. The latter two histidines were those from the neighbouring dimer. The sixth coordination of the Mg²⁺ ion was with a water molecule, which was anchored by a hydrogen bond with the side chain of Glu 81. To substantiate the residues at the metal-ion-binding site on HutP, the residues that interact with the metal ions—His 73, His 77 and His 138—were individually replaced with Ala, and the protein–RNA interactions were measured (Fig. 3b). These assays showed clearly that all three residues in the histidine cluster were important for RNA binding through coordinating with a Mg²⁺ ion. Thus, the Mg²⁺ ion contributes towards mediating the interaction between HutP and L-histidine, along with the residues within the specific binding site in the dimer and the dimer–dimer interface of HutP. However, the Mg²⁺ ion does not bind directly to the RNA. This observation is in contrast with the role of metal ions reported in the past; for example, in the L11–RNA complexes, the metal ions occupy crucial locations to stabilize the sharp turns of the junctions, strengthening the overall structure of the four-way junction²⁴. The overall quaternary complex structure is unique and distinct from those of previous anti-terminator/attenuator complexes^{22,23} and

single-stranded RNA-binding complexes^{25–31}. Nevertheless, the key residues identified in the present study for the interactions with Mg^{2+} (His 73, His 77 and His 138), L-histidine (Tyr 69, Glu 81, Arg 88 and Arg 98) and RNA (Glu 55, Thr 56, Thr 99, Thr 128 and Glu 137) are completely conserved in all HutP proteins from *Bacillus* species (Supplementary Fig. S4).

Structural rearrangements

The overall quaternary complex of HutP resembles the HutP–HBN complex structure (r.m.s.d. 2.45 Å); however, the quaternary complex reveals some structural changes in HutP in comparison with the HutP–HBN complex, especially in the L-histidine-binding site (Fig. 4a) and the loop regions L3 (Fig. 4b), L4 (Supplementary

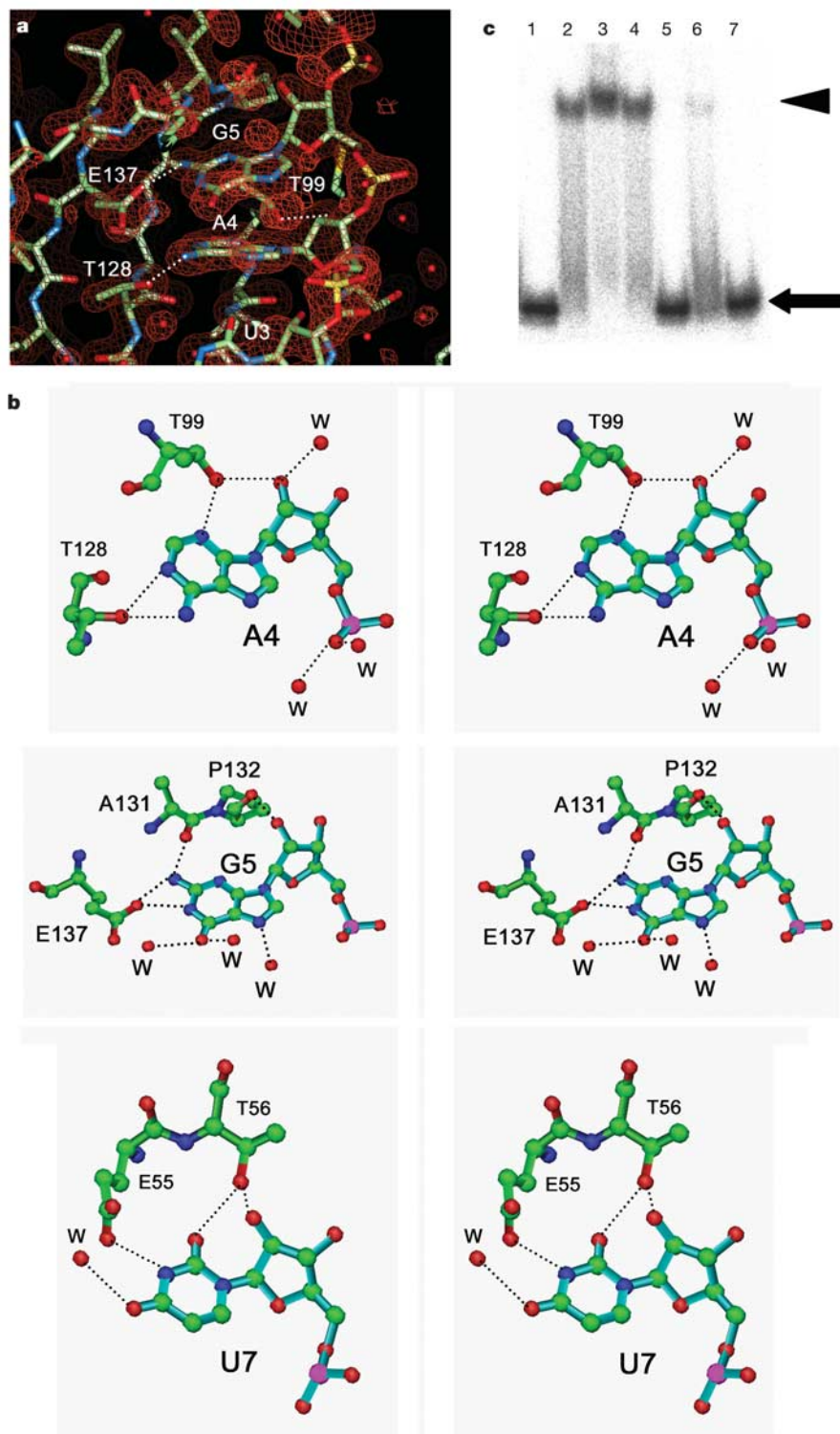


Figure 2 HutP–RNA interactions. **a**, Portion of the electron density map. The protein and RNA are both shown as ball-and-stick models. **b**, A close-up stereo view of the hydrogen-bonding interactions with RNA, shown by broken lines. The protein interactions with the bases and the ribose moieties of A4, G5 and U7 are shown. **c**, Mutational analysis of potential RNA-binding-site residues. The purified protein (1,200 nM) was equilibrated in

HEPES buffer (pH 7.4) and mixed with the 21-mer RNA at 20 nM (labelled RNA 10^4 c.p.m.) in the presence of 10 mM L-histidine. Lane 1, input RNA (21-mer); lane 2, HutP; lane 3, Glu55Ala; lane 4, Thr56Ala; lane 5, Thr99Ala; lane 6, Thr128Ala; lane 7, Glu137Ala. Free RNA and complexed RNA are indicated by an arrow and an arrowhead, respectively.

Fig. S5) and L5 (Fig. 4c), on binding to the L-histidine, metal ions and RNA. It seems that once the Mg^{2+} ion is bound by the HutP–L-histidine complex, the Mg^{2+} ion might pull the L-histidine out of the hydrophobic pocket and bind to it tightly. In this way the Mg^{2+} -bound L-histidine could move along the dimer interface and reside about 12 Å from the pocket. This movement of L-histidine is apparently linked to the movement of Arg88 in the opposite direction, leading to the disruption of the hydrophobic pocket formed by the salt bridge between Arg88 and Glu81 in the HutP–HBN complex, and the formation of a new salt bridge between Arg88 and Arg98 (Fig. 4a). This drastic rearrangement of Arg98 (C_{α} position shifted by 5.4 Å) is accompanied by a large change (5.0 Å) in the C_{α} position of the next residue, Thr99 (Fig. 4b). Consequently, the Thr99 side chain forms two hydrogen bonds with the N3 and 2'-OH of the A4 base. On the other end, the reorientation of the His138 imidazole ring to coordinate with the Mg^{2+} ion causes a large conformational change in the local backbone chain, particularly in the torsion angles around His138 and the next residue, Glu137. As a result, the backbone oxygen of Glu137 bonds to the hydroxyl group of Tyr112. This changes the orientation of the side chain of Glu137, allowing it to interact specifically with G5 through two hydrogen bonds (Fig. 4c). These changes in HutP might have been initiated by the Mg^{2+} ion and transmitted along the backbone chain to Thr128 through the L5 loop. The side chain of Thr128 thus moves to a preferable position to interact specifically with the A4 base through two hydrogen bonds. In addition, the backbone oxygen atoms of Ala131 and Pro132, located within the L5 loop, are involved in hydrogen bonds with the G5 base. Thus, the observed specific interactions between the HutP residues and the conserved RNA residues A4 and G5 within the UAG motif are probably initiated by the Mg^{2+} ion after it becomes bound by the HutP–L-histidine complex.

Structures of HutP and HutP–L-histidine– Mg^{2+}

Although the present quaternary complex structure explains several molecular interactions with ligands, including the importance of the Mg^{2+} ions, which is consistent with our biochemical observation that HutP requires L-histidine and Mg^{2+} ions, it was difficult to pinpoint the specific ligand that causes the structural rearrangement in HutP on the basis of the two crystal structures (quaternary complex and HutP–HBN complex) mentioned above. To identify

the ligand responsible for the structural rearrangement of HutP and to identify the intermediate structures assumed by HutP towards attaining the active anti-terminator complex, we crystallized and solved two additional structures of the HutP protein in the presence of EDTA and in the presence of L-histidine and $MgCl_2$. To probe the structure of HutP (uncomplexed form) in the absence of any Mg^{2+} ions, we used EDTA to chelate the endogenous Mg^{2+} ions that might have been present during purification. These two structures were solved by molecular replacement with the HutP–HBN complex as a search model (PDB ID 1VEA) and were refined to resolutions of 2.80 and 1.70 Å, respectively (Supplementary Table 1).

Comparing the uncomplexed HutP and HutP–L-histidine– Mg^{2+} complex structures with the quaternary complex (HutP–L-histidine– Mg^{2+} –RNA) and our previously determined HutP–HBN complex, the uncomplexed HutP protein shares overall similarity on superposition with the HutP–HBN complex (r.m.s.d. 1.17 Å), and similarly the HutP–L-histidine– Mg^{2+} complex correlates well with the quaternary complex (r.m.s.d. 0.69 Å) (Supplementary Fig. S6a, b). Although the overall HutP–L-histidine– Mg^{2+} complex resembles those of the uncomplexed HutP and the HutP–HBN complex, the local structures, particularly the connecting parts of β -strands and the loop regions L3, L4 and L5, have undergone a significant structural rearrangement, as observed in the quaternary complex. An r.m.s.d. of 3.1 Å was observed between the HutP–L-histidine– Mg^{2+} complex and the uncomplexed HutP, and similarly an r.m.s.d. of 3.15 Å was observed between the HutP–L-histidine– Mg^{2+} complex and the HutP–HBN complex. The structural rearrangements in the β -sheet and loops L3 and L5 are the most important for recognizing the RNA; however, loop L4 is flexible in all of the structures. We have also crystallized and solved the structure of HutP in the presence of Mg^{2+} ions, and the structure is identical to those of the uncomplexed HutP and the HutP–HBN complex (T.S.K., H.M. & P.K.R.K., unpublished data). The uncomplexed HutP (or HutP structure in the presence of Mg^{2+} ion) structure reveals the existence of a hydrophobic pocket similar to that observed in the HutP–HBN complex (Fig. 5a, b). From these revelations it is clear that the existence of a hydrophobic pocket is a necessary prerequisite for the recognition of the L-histidine imidazole group and its discrimination from other amino acid residues. The importance of these interactions was evaluated previously, by

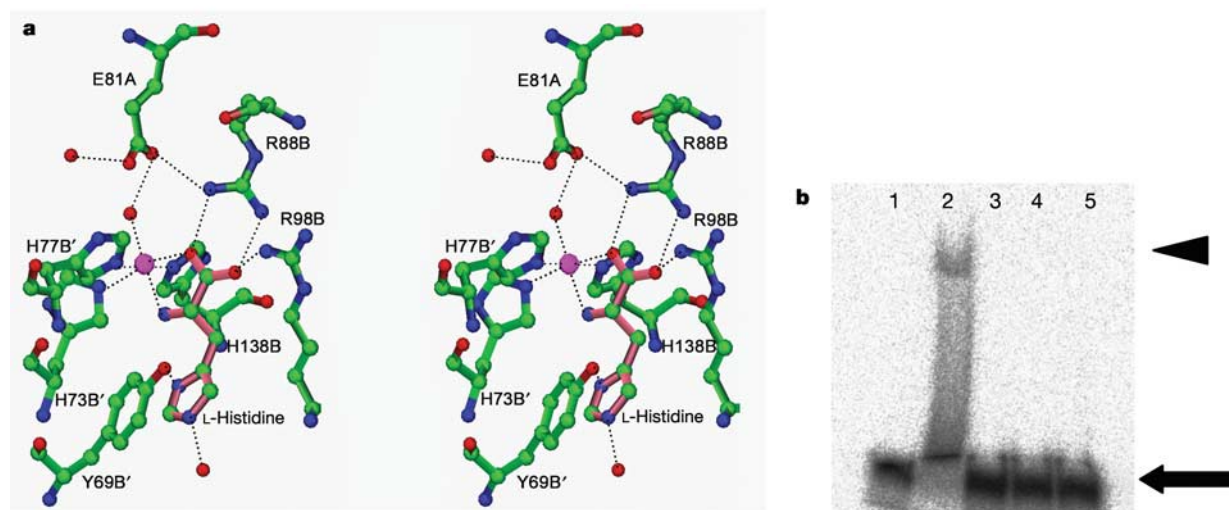


Figure 3 L-histidine and Mg^{2+} interactions in the HutP quaternary complex. **a**, A close-up stereo view of the L-histidine and Mg^{2+} binding site in the HutP quaternary complex. Hydrogen bonds are indicated by dotted lines. L-Histidine is represented by a cyan-coloured 'ball-and-stick' model. The Mg^{2+} and water molecules are represented by cpk models in pink and red, respectively. **b**, Mutational analysis of potential residues at the

metal-ion-binding site. Binding reactions were performed as described for Fig. 2c. Lane 1, input RNA (21-mer); lane 2, HutP; lane 3, His73Ala; lane 4, His77Ala; lane 5, His138Ala. Free RNA and complexed RNA are indicated by an arrow and an arrowhead, respectively.

using L-histidine analogues with modifications on the imidazole group in the L-histidine and mutational analyses of the protein residues that interact with the L-histidine imidazole group^{16,17}.

The HutP–L-histidine–Mg²⁺ complex (Fig. 5c) closely resembles that of the quaternary complex (HutP–L-histidine–Mg²⁺–RNA) of

HutP, even in the absence of RNA (Fig. 5d). This suggests that the RNA is not responsible for the observed structural rearrangement, but the metal ions and the L-histidine ligand cause these changes before recognizing the target RNA. In these complexes, the metal ion and L-histidine interact mutually and facilitate the required

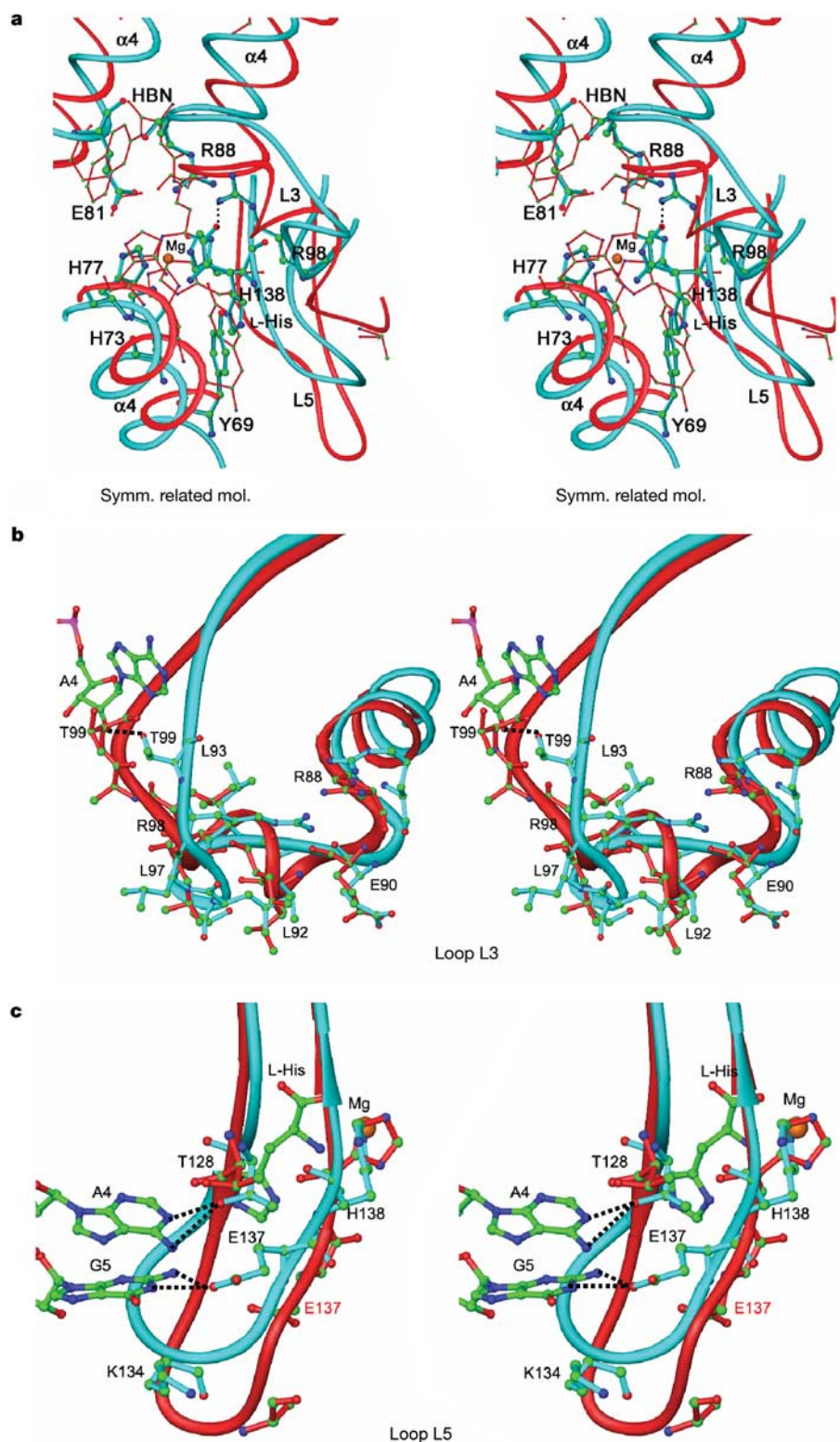


Figure 4 Stereo views of conformational changes observed in the quaternary complex. **a**, A comparison between the quaternary complex (blue) and the HutP–HBN complex (red) showing the structural differences around the L-histidine and Mg²⁺ binding site. Symm.

related mol., neighbouring dimer within the hexamer. **b**, Conformational changes observed in the loop L3 region, coloured as in **a**. **c**, Conformational changes observed in loop L5, coloured as in **a**.

overall structural rearrangement, because if one of the two components (metal ion or L-histidine) is absent, HutP fails to bind to the RNA to form an anti-termination complex. However, L-histidine recognition must be the first step in this process, because until the imidazole ring of the L-histidine ligand is verified in the hydrophobic pocket, the HutP is not ready for the Mg^{2+} -ion-mediated structural rearrangement. At this point, the Mg^{2+} ion is important in bringing about the overall structural rearrangement, the repositioning of L-histidine and the coordination with the HutP residues. However, the Mg^{2+} ions themselves cannot bring about the overall structural rearrangement in the absence of L-histidine, because the L-histidine contributes two coordinations for the Mg^{2+} ion, which might be important to anchor it. The loss of even one coordination to the Mg^{2+} ion prevents quaternary complex formation, as shown by the mutational analysis (H73A, H77A and H138A) of the metal-binding site. When L-histidine is absent, two interactions with the metal ion are lost, in the HutP–L-histidine– Mg^{2+} complex and the quaternary complex, and that is why the metal ion itself might not facilitate the structural rearrangement. From these analyses it is clear that the specific metal ion coordinations with the histidine cluster of HutP and the L-histidine ligand are mandatory for the structural rearrangement.

In considering the various structural and biochemical studies described above, it is tempting to speculate about the structural

rearrangement of HutP at each stage (Fig. 5e). Initially, HutP exists as a hexamer that forms a hydrophobic pocket to attract the L-histidine ligand into the pocket (step A). The next step, the so-called 'tolling step' (step B), verifies the incoming residue in the pocket, to discriminate the imidazole-containing residue from others, and here L-histidine interacts extensively with the HutP residues. This leads to the rearrangement of the local structures near the L-histidine-binding site, such as the one favouring the formation and stabilization of the $\beta 4$ strand. Once the L-histidine interactions have been verified in the pocket, the HutP undergoes a significant structural rearrangement in the presence of Mg^{2+} ions. In this step (step C), the L-histidine moves away from the hydrophobic pocket and establishes new interactions with HutP and Mg^{2+} ions. This changes the L3 and L5 loops into the activated form. Once the HutP has been activated by the Mg^{2+} ion and L-histidine, it recognizes the specific RNA sequence within the terminator region on both sides of the hexamer surface (top and bottom), yielding a triangular shape of RNA, but without undergoing any further structural rearrangement in the protein side (step D).

Although the above structural and biochemical studies indicate that hexameric HutP recognizes a specific RNA sequence within the terminator on either surface, it is intriguing to consider how the terminator structure is being altered. What is the physiological relevance of the existence of two RNA-binding sites in HutP? After

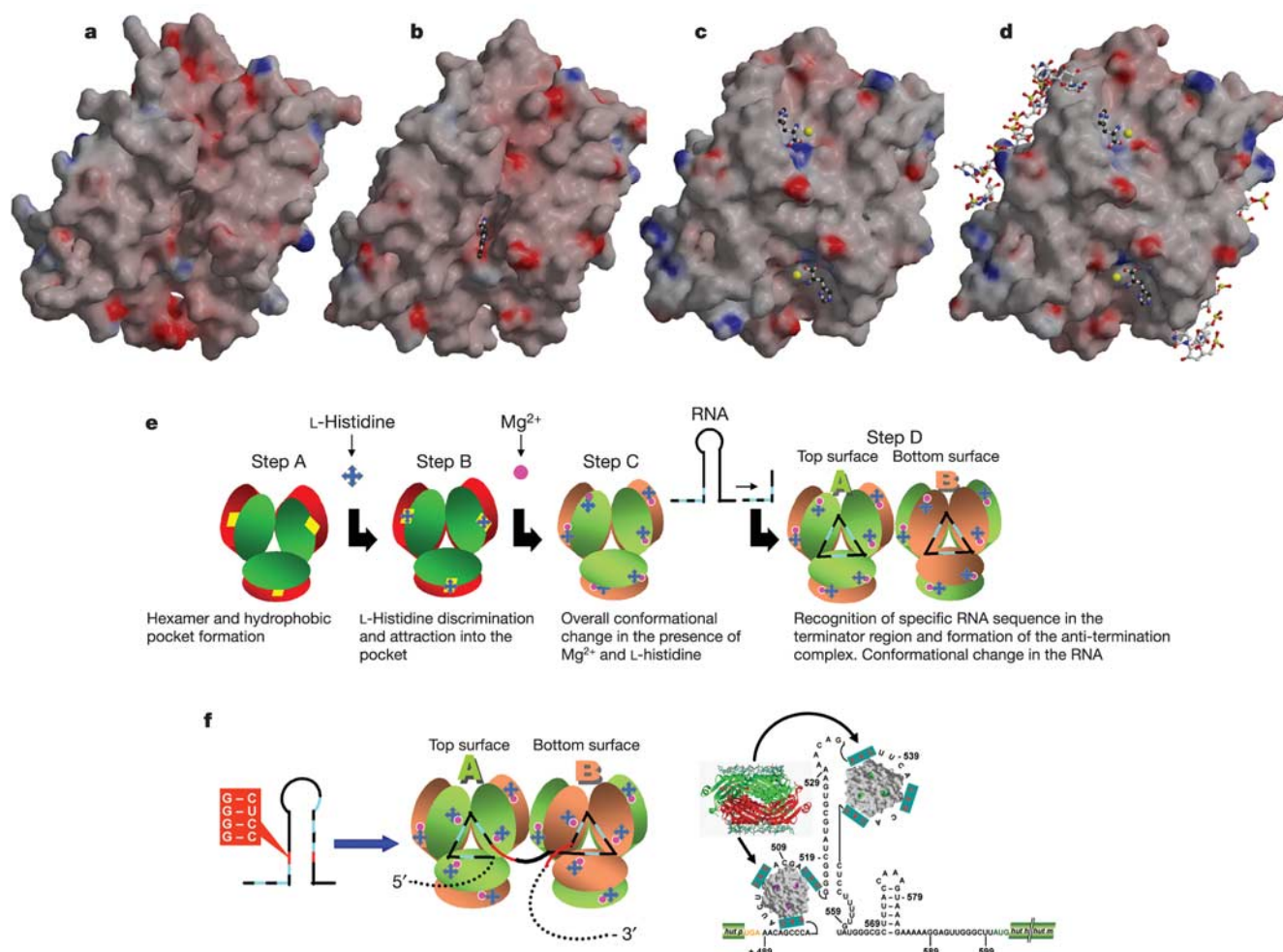


Figure 5 Electrostatic surface potential models of HutP and the proposed mechanism for the anti-terminator complex formation. **a–d**, Molecular surfaces of uncomplexed HutP (**a**), HutP–HBN (**b**), the HutP–L-histidine– Mg^{2+} complex (**c**) and the quaternary HutP complex (**d**), coloured in accordance with the electrostatic potential. HBN, L-histidine and RNA are represented by ball-and-stick models. Mg^{2+} ions are

represented by a cpk model. **e**, A schematic model proposed for HutP anti-terminator complex formation. **f**, A proposed model for the existence of two potential binding sites (highlighted in blue boxes) within the terminator region. The GC-rich region is highlighted in the red box. The RNA-binding residues are indicated by magenta and green.

re-examining the entire intervening sequence (nucleotides +489 to +600, between the stop codon of *hutP* and the initiation codon of the *hutH* gene; Fig. 1a), on the basis of the above evidence we found that there are two potential HutP-binding sites (site I, +498 to +514; site II, +535 to +549). Each binding site consists of three NAG motifs (Fig. 5f). In these two potential binding sites, the first recognizes on one surface of the hexameric HutP and the second recognizes the other surface. Between these two binding sites a 20-nucleotide spacer region exists, which is sufficient to reach the RNA on the other surface of the hexameric HutP. It is also interesting to observe that the HutP binding sites are located in the upstream and downstream regions of the GC-rich region of the terminator (Fig. 5f). We therefore propose that a single *hut* mRNA binds on each side of the hexameric HutP protein and by this interaction it can completely destabilize the terminator structure and allow RNA polymerase to pass through the terminator to synthesize the full-length transcript. Further structural and biochemical analyses are required for an understanding of this mode of HutP–RNA recognition; these studies are currently in progress in our laboratory.

Conclusions

The present structural studies revealed the active structure of the anti-terminator complex (HutP–L-histidine–Mg²⁺–RNA) that regulates the transcription of *hut* mRNA. The structure explains how the HutP and RNA interactions are regulated critically by the L-histidine and the Mg²⁺ ion through structural rearrangements. After analysing the structures of HutP, uncomplexed HutP, the HutP–L-histidine–Mg²⁺ complex and the quaternary complex HutP–L-histidine–Mg²⁺–RNA, we could identify not only the intermediate structures of HutP (before forming an active structure) but also unravel the importance of the interactions of the Mg²⁺ ion in the complexes. □

Methods

Protein and RNA preparation

HutP was overexpressed and purified from *Escherichia coli* as described previously³². The purified protein was concentrated to 9 mg ml^{−1} and equilibrated with binding buffer (15 mM HEPES pH 7.4, 30 mM NaCl, 5 mM MgCl₂). The 21-mer RNA oligonucleotides (wild type, 5′-CAUAGAUCUAGACGAUAGGG-3′; 4-U spacer, 5′-UUUAGUUUUUAGUUUUUAGUU-3′) and shorter RNAs (5′-UAG-3′, 5′-UUAG-3′, 5′-UAGU-3′ and 5′-UUUAGUU-3′) were synthesized chemically and subsequently deprotected, purified by denaturing PAGE and annealed as described previously¹⁸.

Crystallization and data collection

Crystals of the HutP–L-histidine–Mg²⁺–21-mer RNA complex were grown at 20 °C by the hanging-drop vapour-diffusion method. Drops consisted of 2 µl protein, 1 µl 0.1 M L-histidine, 1 µl RNA and 2 µl well solution (100 mM HEPES pH 7.4, 0.1 M MgCl₂, 40% MPD (methyl-2,4-pentanediol). The molar ratio of protein to RNA used for crystallization was 1:2. Well-ordered crystals were grown within a week with the 4U spacer RNA. To crystallize the HutP protein without metal ions (uncomplexed HutP), we used EDTA as the chelating agent. The uncomplexed HutP crystals were grown from drops consisting of 2 µl protein, 1 µl 0.1 M EDTA and 2 µl well solution (100 mM Bicine pH 7.6, 30% PEG550MME). HutP–L-histidine–Mg²⁺ crystals were grown from drops consisting of 2 µl protein, 1 µl 0.1 M L-histidine and 2 µl well solution (100 mM HEPES pH 7.4, 0.1 M MgCl₂, 40% MPD). Diffraction data were collected with a charge-coupled device (ADSC) detector at the Photon Factory, Tsukuba, Japan. Data sets for the quaternary, uncomplexed HutP and HutP–L-histidine–Mg²⁺ structures were processed at 1.60, 2.78 and 1.70 Å, respectively, with the HKL2000 suite of programs³³. The quaternary, uncomplexed HutP and HutP–L-histidine–Mg²⁺ crystals belonged to the R3, P2₁3 and P2₁2₁2 space groups, respectively (Supplementary Table 1).

Structure determination

All crystallization trials were performed with HutP, L-histidine, MgCl₂ and various synthetic RNAs containing different lengths and sequences. Crystallization trials of the quaternary complex with wild-type RNA (5′-CAUAGAUCUAGACGAUAGGG-3′) or with shorter RNAs (5′-UAG-3′, 5′-UUAG-3′, 5′-UAGU-3′ and 5′-UUUAGUU-3′) did not yield crystals. The crystals were obtained with a 21-mer RNA, 5′-UUUAGUUUUUAGUUUUUAGUU-3′, which possesses higher affinity over the wild-type sequence and conserves all of the bases important for HutP recognition within the *hut* mRNA terminator region. The HutP quaternary complex (HutP–L-histidine–Mg²⁺–21-mer RNA), uncomplexed HutP and HutP–L-histidine–Mg²⁺ complex structures were determined by molecular replacement, using the HutP–HBN complex as a search model

(PDB ID 1VEA). Both molecular replacement and model refinement were performed with CNS³⁴. The protein and RNA model was built with the program Quanta³⁵. The structures of the HutP quaternary, uncomplexed HutP and HutP–L-histidine–Mg²⁺ complexes were refined to 1.60, 2.80 and 1.70 Å, respectively (Supplementary Table 1). The crystallographic asymmetric unit of the quaternary complex model contains two monomers (A and B) of HutP (except for residues 21–23 of molecule A and residues 21–24 of molecule B), two Mg²⁺ ions, two L-histidines, two seven-nucleotide fragments from the 21-mer RNA and 314 water molecules in the asymmetric unit. Residues 1–5 of molecule A and 1–4 of molecule B in the final model of uncomplexed HutP, and the N-terminal residues of the A, B and C molecules in the HutP–L-histidine–Mg²⁺ complex, were not modelled because of the absence of electron density. Also, A28Glu of the quaternary complex, A6Glu and B5Lys of uncomplexed HutP, and A5Lys, A19Asn, A20Glu, A21Glu, A23Glu, A24Ser, A28Glu, A41Lys, A134Lys, B5Lys, B22Glu, B23Glu, B24Ser, B28Glu, B115Glu, C5Lys, C22Glu, C23Glu, C24Ser, C26Gln, C28Glu, C29Glu, C32Arg, C41Lys, C60Lys, C67Glu and C115Glu of the HutP–L-histidine–Mg²⁺ complex were modelled with Ala, because the side-chain densities were lacking for these residues. Figures were prepared with programs Quanta³⁵, Ribbons³⁶, Grasp³⁷, Molscript³⁸ and Raster3D³⁹.

Mutagenesis of HutP, gel mobility shift and filter binding assays

The important residues for RNA and metal ion interactions on the protein surface, Glu 55, Thr 56, Thr 99, Thr 128, Glu 137, His 73, His 77 and His 138, were each replaced by Ala, and the mutant proteins were purified as described previously¹⁶. The *hut* 21-mer RNA was synthesized chemically, 5′ end labelled with ³²P and purified. We performed three independent experiments with three different protein concentrations (300, 600 and 1,200 nM). The reaction mixtures of the quaternary complex were prepared in binding buffer and resolved by 10% native PAGE as described previously^{16,18}. To determine the equilibrium dissociation constant (*K*_d) we performed a filter binding assay as described previously^{16,18}.

Received 11 October 2004; accepted 10 January 2005; doi:10.1038/nature03355.

- Chasin, L. A. & Magasanik, B. Induction and repression of the histidine-degrading enzymes of *Bacillus subtilis*. *J. Biol. Chem.* **243**, 5165–5178 (1968).
- Kimmhi, Y. & Magasanik, B. Genetic basis of histidine degradation in *Bacillus subtilis*. *J. Biol. Chem.* **245**, 3545–3548 (1970).
- Oda, M., Sugishita, A. & Furukawa, K. Cloning and nucleotide sequences of histidase and regulatory genes in the *Bacillus subtilis* *hut* operon and positive regulation of the operon. *J. Bacteriol.* **170**, 3199–3205 (1988).
- Wray, L. V. Jr & Fisher, S. H. Analysis of *Bacillus subtilis* *hut* operon expression indicates that histidine-dependent induction is mediated primarily by transcriptional anti-termination and that amino acid repression is mediated by two mechanisms: regulation of transcription initiation and inhibition of histidine transport. *J. Bacteriol.* **176**, 5466–5473 (1994).
- Oda, M., Kobayashi, N., Ito, A., Kurusu, Y. & Taira, K. *Cis*-acting regulatory sequences for anti-termination in the transcript of the *Bacillus subtilis* *hut* operon and histidine-dependent binding of HutP to the transcript containing the regulatory sequences. *Mol. Microbiol.* **35**, 1244–1254 (2000).
- Yanofsky, C. Advancing our knowledge in biochemistry, genetics, and microbiology through studies on tryptophan metabolism. *Annu. Rev. Biochem.* **70**, 1–37 (2001).
- Gollnick, P. & Babitzke, P. Transcription attenuation. *Biochim. Biophys. Acta* **1577**, 240–245 (2002).
- Stulke, J. Control of transcription termination in bacteria by RNA-binding proteins that modulate RNA structures. *Arch. Microbiol.* **177**, 433–440 (2002).
- Houman, E., Diaz-Torres, M. R. & Wright, A. Transcriptional anti-termination in the *bgl* operon of *E. coli* is modulated by a specific RNA binding protein. *Cell* **62**, 1153–1163 (1990).
- Aymerich, S. & Stenmetz, M. Specificity determinants and structural features in the RNA target of the bacterial anti-terminator proteins of the BglG/SacY family. *Proc. Natl Acad. Sci. USA* **89**, 10410–10414 (1992).
- Babitzke, P. & Yanofsky, C. Reconstitution of *Bacillus subtilis* trp attenuation *in vitro* with TRAP, the trp RNA-binding attenuation protein. *Proc. Natl Acad. Sci. USA* **90**, 133–137 (1993).
- Arnaud, M. D., Debarbouille, M., Rapoport, G., Saier, M. H. Jr & Reizer, J. *In vitro* reconstitution of transcriptional attenuation by the SacT and SacY proteins of *Bacillus subtilis*. *J. Biol. Chem.* **271**, 18966–18972 (1996).
- Lu, Y., Turner, R. J. & Switzer, R. L. Function of RNA secondary structures in transcriptional attenuation of the *Bacillus subtilis* *pyr* operon. *Proc. Natl Acad. Sci. USA* **93**, 14462–14467 (1996).
- Alpert, C. A. & Siebers, U. The *lac* operon of *Lactobacillus casei* contains *lacT*, a gene coding for a protein of the BglG family of transcriptional anti-terminators. *J. Bacteriol.* **179**, 1555–1562 (1997).
- Glatz, E., Nilsson, R. P., Rutberg, L. & Rutberg, B. A dual role for the *Bacillus subtilis* glpD leader and the GlpP protein in the regulated expression of glpD: anti-termination and control of mRNA stability. *Mol. Microbiol.* **19**, 319–328 (1996).
- Kumarevel, T. S. *et al.* Crystal structure of activated HutP: an RNA binding protein that regulates hut operon in *Bacillus subtilis*. *Structure* **12**, 1269–1280 (2004).
- Kumarevel, T. S., Mizuno, H. & Kumar, P. K. R. Allosteric activation of HutP protein, that regulates transcription of hut operon in *Bacillus subtilis*, mediated by various analogs of histidine. *Nucleic Acids Res. Suppl.* **3**, 199–200 (2003).
- Kumarevel, T. S., Gopinath, S. C. B., Mizuno, H. & Kumar, P. K. R. Identification of important chemical groups of the *hut* mRNA for HutP interactions that regulates the *hut* operon in *Bacillus subtilis*. *Nucleic Acids Res.* **32**, 3904–3912 (2004).
- Read, T. D. *et al.* The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423**, 81–86 (2003).
- Ivanova, N. *et al.* Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature* **423**, 87–91 (2003).
- Takami, H. *et al.* Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res.* **28**, 4317–4331 (2000).
- Antson, A. A. *et al.* Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature* **401**, 235–242 (1999).
- Yang, Y., Declercq, N., Manivel, X., Aymerich, S. & Kochoyan, M. Solution structure of the LicT-RNA anti-termination complex: CAT clamping RAT. *EMBO J.* **21**, 1987–1997 (2002).

24. Wimberly, B. T., Guymon, R., McCutcheon, J. P., White, S. W. & Ramakrishnan, V. A detailed view of a ribosomal active site: The structure of the L11–RNA complex. *Cell* **97**, 491–502 (1999).
25. Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burely, S. K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98**, 835–845 (1999).
26. Handa, N. *et al.* Structural basis for recognition of the *tra* mRNA precursor by the Sex-lethal protein. *Nature* **398**, 579–585 (1999).
27. Liu, Z. *et al.* Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294**, 1098–1102 (2001).
28. Lewis, H. A. *et al.* Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and fragile X syndrome. *Cell* **100**, 323–332 (2000).
29. Bogden, C. E., Fass, D., Bergman, N., Nichols, M. D. & Berger, J. M. Structural basis for terminator recognition by the Rho transcription termination factor. *Mol. Cell* **3**, 487–493 (1999).
30. Thore, S., Mayes, C., Sauter, C., Weeks, S. & Suck, D. Crystal structures of the *Pyrococcus abyssi* Sm core and its complex with RNA. Common features of RNA binding in Archaea and Eukarya. *J. Biol. Chem.* **278**, 1239–1247 (2003).
31. Nagai, K., Outbridge, C., Jessen, T. H., Li, J. & Evans, P. R. Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. *Nature* **348**, 515–520 (1990).
32. Kumarevel, T. S. *et al.* Crystallization and preliminary X-ray diffraction studies of HutP protein: an RNA binding protein that regulates the transcription of *hut* operon in *Bacillus subtilis*. *J. Struct. Biol.* **138**, 237–240 (2002).
33. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation model. *Methods Enzymol.* **276**, 307–326 (1997).
34. Brunger, A. *et al.* Crystallography and NMR system: a new software system for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
35. Oldfield, T. J. A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent. *Acta Crystallogr. D* **57**, 82–94 (2001).
36. Carson, M. Ribbons. *Methods Enzymol.* **277**, 493–505 (1997).
37. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).
38. Kraulis, P. J. Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
39. Merritt, E. A. & Bacon, D. J. Raster3D photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the personnel at beamline AR-NW12 of the Photon Factory, Tsukuba, Japan for assistance during data collection, and T. Misono for help in the purification of HutP mutants and in sequence analysis. T.S.K. holds a fellowship of the AIST, Tsukuba, Japan. This work was supported by funds from METI and the ORCS project to P.K.R.K. and H.M., respectively.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.K.R.K. (pkk-kumar@aist.go.jp). Coordinates for the HutP quaternary complex, uncomplexed HutP and the HutP–L-histidine–Mg²⁺ complex have been deposited in the Protein Data Bank under accession codes 1WMQ, 1WPS and 1WPV, respectively.

An upper limit to the masses of stars

Donald F. Figer

STScI, 3700 San Martin Drive, Baltimore, Maryland 21218, USA

There is no accepted upper mass limit for stars. Such a basic quantity eludes both theory and observation, because of an imperfect understanding of the star-formation process and because of incompleteness in surveying the Galaxy¹. The Arches cluster^{2–7} is ideal for investigating such limits, being large enough to expect stars at least as massive as ~ 500 solar masses ($\sim 500 M_{\odot}$; based on a typical mass function), and young enough for its most massive members to still be visible. It is also old enough to be free of its natal molecular cloud, it is at a well-established distance, and it is close enough for us to discern its individual stars². Here I report an absence of stars with initial masses greater than $130 M_{\odot}$ in the Arches cluster, whereas the typical mass function predicts 18. I conclude that this indicates a firm limit of $150 M_{\odot}$ for stars; the probability that the observations are consistent with there being no upper limit is 10^{-8} .

Theory provides little guide in determining the most massive star that can form. Pulsational instabilities were once thought to destroy stars more massive than $95 M_{\odot}$ (ref. 8); however, these pulsations may be damped⁹. Radiation pressure, and/or ionizing flux, inhibit accretion for stellar masses greater than $60 M_{\odot}$ (ref. 10), but direct collisions of protostellar clumps may overcome these effects¹¹. Although stellar evolution models have been computed for massive stars covering a large range in mass, up to $1,000 M_{\odot}$ (refs 12, 13), no such stars have ever been observed. Indeed, some of the most massive candidates have proved to be systems of multiple stars¹⁴.

Stars generally form with a frequency that decreases with increasing mass for masses greater than $\sim 1 M_{\odot}$, that is, $d(\log N)/d(\log m) = \Gamma$, where m is the initial stellar mass, N is the number of stars per logarithmic initial mass interval, and Γ is observed to be -1.35 (refs 15, 16). For stellar clusters young enough not to have lost members to supernovae, the distribution of stars is populated to the point where the mass function predicts one star, within the uncertainties of low number statistics. Therefore, stars with mass $M > 150 M_{\odot}$ can only be observed in very massive clusters with total stellar mass $> 10^4 M_{\odot}$. This requirement limits the potential sample of stellar clusters that can constrain the upper mass limit. Only a few clusters in the Galaxy satisfy this requirement, and all are located in the Galactic Centre.

To investigate the possibility that stars with $M > 150 M_{\odot}$ exist, imaging data were obtained using the Near-Infrared Camera and Multi-Object Spectrometer instrument on the Hubble Space Telescope in a programme to measure the mass functions of the most massive young clusters in the Galaxy, near the Galactic Centre². Intervening dust prevents observations of these clusters at optical or ultraviolet wavelengths, so images were obtained in near-infrared wavelengths (see Supplementary Fig. 1). Nearby control fields were also imaged to estimate the number of field stars that contaminate observations in such a densely populated region.

Photometry for stellar sources in the images was extracted, and corrected for the absorbing effects of dust by comparing the observed colours to those expected for the appropriate spectral types; note that intrinsic colours of massive stars on the main sequence at infrared wavelengths differ by only a few per cent. The dereddened fluxes were then converted into bolometric fluxes by accounting for the distance to the Galactic Centre, and the Geneva stellar evolution models were used to infer initial masses for each star¹⁷ (see Fig. 1). Although these models have associated errors,

note that the Arches stars are relatively unevolved; indeed, only the brightest dozen or so members show evidence of chemical enrichment by nucleosynthetic processes¹⁸. Some of the brightest stars in the cluster (three to ten, depending on cluster age within a range of 2–2.5 Myr and the coefficients in the extrapolation law) extend just above the $120 M_{\odot}$ limit of the mass–flux relation; I estimate masses for them that do not exceed $130 M_{\odot}$ through an extrapolation of this relation (see Supplementary Fig. 2).

The initial masses I estimate here agree with those inferred through wind/atmosphere modelling of high-resolution spectral observations to within a few per cent (ref. 3). Others have also applied the same technique to construct mass functions from infrared observations of massive young clusters, showing that these determinations are consistent with those estimated from optical observations¹⁹. In addition, several groups have found good consistency in physical properties inferred from optical and infrared analyses for massive stars at all stages of evolution^{20–22}.

Figure 2 shows the resultant initial mass function of the Arches cluster, assuming an age of 2 Myr, for stars within a projected radius of 0.5 pc, and solar metallicity^{2,18}. Although the cluster is the densest

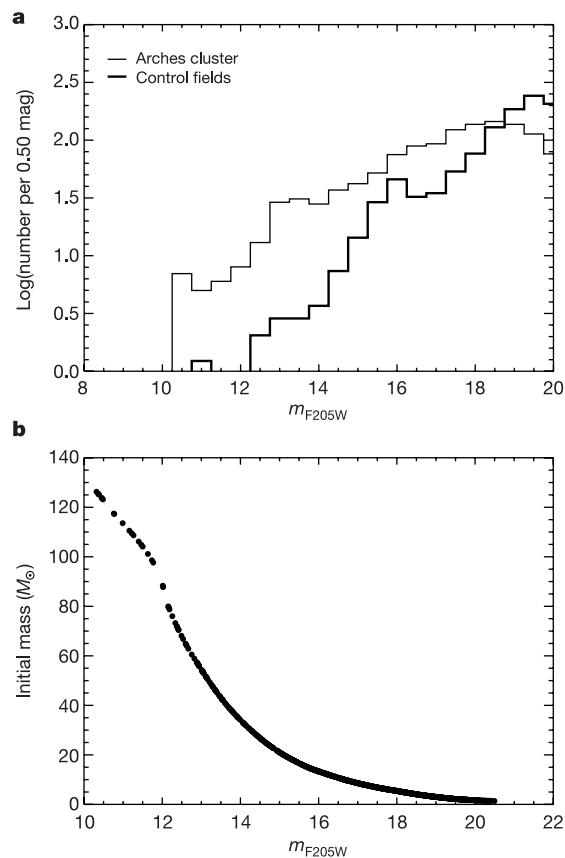


Figure 1 Observed frequency distribution and inferred masses of stars in the Arches cluster versus brightness. Here $m_{F205W} = -2.5 \log(F/F_{\text{Vega}})$, where F is the flux of the star and F_{Vega} is the flux of Vega, as measured in the F205W ($\lambda_{\text{centre}} = 2.05 \mu\text{m}$) filter in NICMOS. **a**, Near-infrared luminosity functions of the central parsec of the Arches cluster (thin line) and nearby background fields (thick line). There are generally fewer bright than faint stars in both fields; but for the vast majority of the brightness range, there are more stars in the Arches cluster than in the control fields. This allows an accurate subtraction of background stars in order to create a mass function for the cluster. The shapes of the distributions are consistent with a very young stellar cluster in the cluster field and an old population ($\tau > 1$ Gyr) in the control fields. **b**, Inferred initial masses for Arches stars, based on the Geneva models¹⁷ for solar abundances and an age of 2 Myr. Each point represents one star in the cluster field. The three brightest stars have masses that slightly exceed $120 M_{\odot}$, the upper limit of the models, and are assigned masses through a linear extrapolation of the mass–flux relation from points immediately below this value.

in the Galaxy³, the data do not suffer from incompleteness due to crowding or sensitivity for the four highest mass bins in the figure. The small amount of background contamination was removed by subtracting the number of stars observed in nearby fields; this resulted in the subtraction of a total of seven stars from the upper four populated mass bins. The frequency distribution generally decreases with increasing mass, and is fitted by two lines through the four most massive populated bins, which contain 39 stars. One line has a slope of $\Gamma = -0.9$, appropriate for the most recent determinations^{2,23}, and the other has a slope equal to the Salpeter value that is observed for most clusters. For both slopes, there appears to be a deficit of expected very massive stars with masses beyond $\sim 130 M_{\odot}$; variations in assumed age (± 0.5 Myr), mass-loss rates and metallicity do not change the result. I estimate cumulative errors of $\sim 10\%$, and conclude conservatively that there is an upper mass cut-off of $\sim 150 M_{\odot}$ (see Supplementary Fig. 3 for the effects of mass loss on the most massive stars).

The observed deficit of stars is significant. If there is no upper mass cut-off, then the odds of identifying no stars beyond the observed limit are 10^{-8} if 18 are expected, and 10^{-14} if 33 are expected, assuming Poisson statistics. In addition, the maximum predicted stellar mass is at least ~ 500 – $1,100 M_{\odot}$, values that are far beyond the masses inferred from the observations. I performed a Monte Carlo simulation of model systems to predict (as a function of cut-off mass) the probability that a cluster with the mass of the Arches cluster could have no stars with initial masses greater than $130 M_{\odot}$ (see Supplementary Fig. 4). In this simulation, I added uncertainties due to differential extinction, photometric error, average cluster age, a spread of ages for individual stars, and error in estimating the average cluster age. The simulation predicts few systems with no stars having initial masses greater than $130 M_{\odot}$ for cut-offs of $150 M_{\odot}$ or greater (Supplementary Fig. 4).

Clearly, the cluster age (τ) is an important quantity for the analysis. If the cluster is too old, $\tau > 3$ Myr, then its most massive

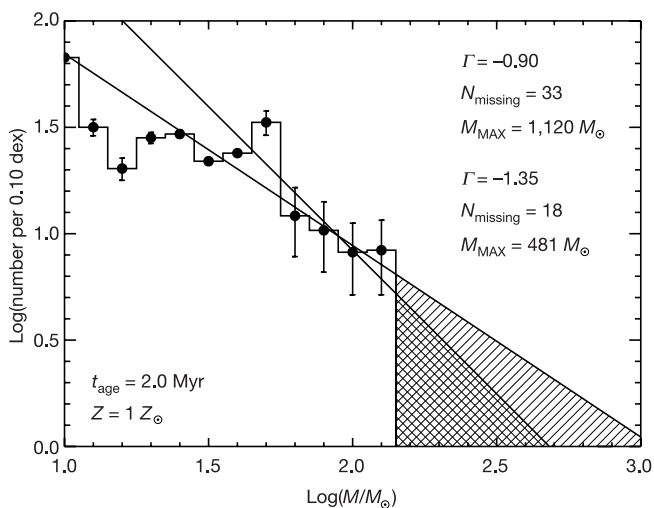


Figure 2 Frequency distribution versus mass for stars in the Arches cluster extracted from data in Fig. 1. The counts in each bin have been reduced by counts in nearby background fields. Error bars represent the Poisson errors based on the background subtracted counts. Two lines are drawn through the average counts in the four highest populated mass bins, with slopes inferred from the data ($d(\log N)/d(\log m) = \Gamma = -0.9$; ref. 2, 23) and that of Salpeter ($\Gamma = -1.35$; ref. 15). For both lines, there is a clear deficit of stars with initial masses greater than $\sim 150 M_{\odot}$, as seen in the hatched regions. In addition, both slopes predict that at least one star in the cluster should have a mass far beyond that observed if there is no upper mass cut-off. N_{missing} is the difference between the number of stars expected with $M_{\text{initial}} > 130 M_{\odot}$ and the number observed. M_{MAX} is the initial mass at which the mass function predicts the existence of one star. t_{age} is the assumed cluster age. Z is the assumed abundance of metals in the cluster stars.

members would no longer be visible (that is, they would have progressed to supernovae), and the observations would then simply reveal an apparent cut-off due to the natural effects of stellar evolution. If the cluster is too young, $\tau \approx 1$ Myr, then the models would predict much higher initial masses for the brightest members; however, note that even younger ages would still require a firm upper mass cut-off, albeit at somewhat higher masses than predicted by the best estimated age. Analyses indicate that the cluster has an age of 2–2.5 Myr (refs 2, 3, 18). A younger age is inconsistent with the nitrogen-enriched atmospheres revealed in the spectra of the most massive stars in the cluster¹⁸. The fairly narrow age range is required by the observed heavy nitrogen enrichment in the brightest stars together with relatively weak observed nitrogen content in the atmospheres of slightly lower mass stars^{3,18}. An older age is inconsistent with the evolutionary status of the most massive stars in the cluster—they have not evolved to advanced stages, such as the carbon Wolf-Rayet phase^{3,6}. In addition, the lack of any supernova remnants in the cluster argues for an age less than 3 Myr. Indeed, if massive stars filling the apparent deficit were formed and evolved to supernovae, one would expect that a supernova remnant would have been formed at least every 50,000 yr for the past 0.5 Myr—yet none are observed. In summary, stars with masses above $\sim 150 M_{\odot}$ should still be visible if they were formed, given my estimate of the age for the Arches cluster.

The observed upper mass limit is on the low side of the estimated masses of a few massive stars in the Galaxy, although it still falls within the error bars of these estimates. It is important to note the large errors in such estimates. For instance, many of these estimates rely on stellar wind/atmosphere models that do not model the effects of increased opacity produced by metals in stellar winds (that is, line-blanketing). With more modern models, new mass estimates are smaller by up to a factor of two. In addition, mass estimates often suffer from uncertainties in distance, reddening and photometry. The typical build-up of errors can easily result in an uncertainty of a factor of two in flux, and of a similar factor in mass estimate. As an example, consider Pismis 24-1, which is estimated to have a mass of 210–290 $M_{\odot}$ (ref. 24). The build-up in errors for this star, from effects described above, produces at least a factor of two variation in flux estimates, and the original mass estimates were produced without the use of line-blanketing. Once these combined effects are included, the true mass of this star may well be below 100 $M_{\odot}$. Note that uncertainty in distance is the next obstacle to making accurate mass estimates once line-blanketing is included; however, the distance to the Galactic Centre is very well known (to within 6%), and the Arches cluster is physically connected to phenomena known to be produced in the Galactic Centre³.

If there are stellar systems more massive than the limit, then perhaps they are binaries, or products of mergers of lower-mass stars. Indeed, the Pistol star, with an inferred initial mass of ~ 150 – $250 M_{\odot}$ (ref. 13), is surrounded by Wolf-Rayet and red supergiant stars that are older than the expected lifetime of such a star²⁵. This apparent paradox may be reconciled if the star is actually multiple, or if it has recently experienced a rejuvenation through a merger with another star²⁶. High-spatial-resolution imaging suggests that the Pistol star is not binary to within a limit of 110 AU (ref. 13), yet massive binaries can have components with orbits on yet-smaller scales¹⁴.

An upper mass cut-off of $\sim 150 M_{\odot}$ was found for the cluster R136 in the low-metallicity environment of the nearby galaxy, the Large Magellanic Cloud²⁷. This result relies on an apparent deficit of ten stars with masses beyond this limit, based on the assumption that R136 has a total stellar mass of $5 \times 10^4 M_{\odot}$; however, this high cluster mass includes stars that span a range of ages, including those that exceed the age at which a massive star is expected to evolve to become a supernova. This has the effect of increasing the base of lower-mass stars from which to extrapolate an expected number of higher-mass stars, thus inflating an apparent deficit if those stars are

not seen. Using a lower estimate of the cluster mass, $2 \times 10^4 M_{\odot}$ (ref. 28), in stars sufficiently young for the present analysis, I estimate that the true deficit beyond $150 M_{\odot}$ in R136 is roughly four stars—that is, the result in the present work is more statistically significant by this measure. If the deficit of massive stars in R136 is real, then it represents another measurement of the upper mass cut-off.

Surprisingly, the cut-off may be similar in environments that span a factor of three in metallicity^{18,29,30}, although metal content is often cited as a proxy for the source of opacity that limits the infall of material and eventual build-up of massive stars. This result implies that the process that limits the mass of a star is independent of metallicity, at least in the range of metallicities primarily found within the Galaxy and the nearby Large Magellanic Cloud. □

Received 5 October; accepted 9 December 2004; doi:10.1038/nature03293.

1. Elmegreen, B. G. Modeling a high-mass turn-down in the stellar initial mass function. *Astrophys. J.* **539**, 342–351 (2000).
2. Figer, D. F. *et al.* HST/NICMOS observations of massive stellar clusters near the Galactic center. *Astrophys. J.* **525**, 750–758 (1999).
3. Figer, D. F. *et al.* Massive stars in the Arches cluster. *Astrophys. J.* **581**, 258–275 (2002).
4. Nagata, T., Woodward, C. E., Shure, M. & Kobayashi, N. Object 17: Another cluster of emission-line stars near the Galactic center. *Astron. J.* **109**, 1676–1681 (1995).
5. Cotera, A. S. *et al.* A new cluster of hot stars near the Galactic center. *Bull. Am. Astron. Soc.* **24**, 1262 (1992).
6. Blum, R. D. *et al.* 2 micron narrowband adaptive optics imaging in the Arches cluster. *Astron. J.* **122**, 1875–1887 (2001).
7. Serabyn, E., Shupe, D. & Figer, D. F. An extraordinary cluster of massive stars near the centre of the Milky Way. *Nature* **394**, 448–451 (1998).
8. Schwarzschild, M. & Harm, R. On the maximum mass of stable stars. *Astrophys. J.* **129**, 637–646 (1959).
9. Beech, M. & Mitalas, R. Formation and evolution of massive stars. *Astrophys. J. Suppl.* **95**, 517–534 (1994).
10. Wolfire, M. G. & Cassinelli, J. P. Conditions for the formation of massive stars. *Astrophys. J.* **319**, 850–867 (1987).
11. Bonnell, I. A., Bate, M. R. & Zinnecker, H. On the formation of massive stars. *Mon. Not. R. Astron. Soc.* **298**, 93–102 (1998).
12. Bond, J. R., Arnett, W. D. & Carr, B. J. The evolution and fate of very massive objects. *Astrophys. J.* **280**, 825–847 (1984).
13. Figer, D. F. *et al.* The Pistol star. *Astrophys. J.* **506**, 384–404 (1998).
14. Daminieli, A. *et al.* η Carinae: binarity confirmed. *Astrophys. J.* **528**, L101–L104 (2000).
15. Salpeter, E. E. The luminosity function and stellar evolution. *Astrophys. J.* **121**, 161–167 (1955).
16. Kroupa, P. The initial mass function of stars: evidence for uniformity in variable systems. *Science* **295**, 82–91 (2002).
17. Schaller, G., Schaerer, D., Meynet, G. & Maeder, A. New grids of stellar models from 0.8 to 120 solar masses at $Z = 0.020$ and $Z = 0.001$. *Astron. Astrophys. Suppl.* **96**, 269–331 (1992).
18. Najarro, F., Figer, D. F., Hillier, D. J. & Kudritzki, R. P. Metallicity in the Galactic center: The Arches cluster. *Astrophys. J.* **611**, L105–L108 (2004).
19. Stolte, A., Brandner, W., Brandl, B., Zinnecker, H. & Grebel, E. K. The secrets of the nearest starburst cluster. I. Very Large Telescope/ISAAC photometry of NGC 3603. *Astron. J.* **128**, 765–786 (2004).
20. Bohannan, B. & Crowther, P. Quantitative near-infrared spectroscopy of OF and WNL stars. *Astrophys. J.* **511**, 374–388 (1999).
21. Hillier, D. J., Crowther, P. A., Najarro, F. & Fullerton, A. W. An optical and near-IR spectroscopic study of the extreme P Cygni-type supergiant HDE 316285. *Astron. Astrophys.* **340**, 483–496 (1998).
22. Najarro, F. Massive stars in the galactic center. *N. Astron. Rev.* **44**, 213–220 (2000).
23. Stolte, A., Grebel, E. K., Brandner, W. & Figer, D. F. The mass function of the Arches cluster from Gemini adaptive optics data. *Astron. Astrophys.* **394**, 459–478 (2002).
24. Walborn, N. R. *et al.* A new spectral classification system for the earliest O stars: Definition of type O2. *Astron. J.* **123**, 2754–2771 (2002).
25. Figer, D. F., McLean, I. S. & Morris, M. Massive stars in the Quintuplet cluster. *Astrophys. J.* **514**, 202–220 (1999).
26. Figer, D. F. & Kim, S. S. in *Stellar Collisions, Mergers and their Consequences* (ed. Shara, M.) 287–295 (ASP Conf. Ser. 263, Astronomical Society of the Pacific, San Francisco, 2002).
27. Weidner, C. & Kroupa, P. Evidence for a fundamental stellar upper mass limit from clustered star formation. *Mon. Not. R. Astron. Soc.* **348**, 187–191 (2003).
28. Walborn, N. R., Maiz-Apellaniz, J. & Barba, R. H. Further insights into the structure of 30 Doradus from the Hubble Space Telescope instruments. *Astron. J.* **124**, 1601–1624 (2002).
29. Ramirez, S. V. *et al.* Stellar iron abundances at the Galactic center. *Astrophys. J.* **537**, 205–220 (2000).
30. de Boer, K. S., Fitzpatrick, E. L. & Savage, B. D. Abundances of O, Mg, S, Cr, Mn, Ti, Ni and Zn from absorption lines of neutral gas in the Large Magellanic Cloud in front of R136. *Mon. Not. R. Astron. Soc.* **217**, 115–126 (1985).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements I acknowledge discussions with P. Najarro, R. Larson, N. Walborn, J. Puls, N. Panagia, M. Morris, C. Weidner, P. Kroupa, R. M. Rich, V. Bromm and M. Livio.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to the author (figer@stsci.edu).

General observation of n-type field-effect behaviour in organic semiconductors

Lay-Lay Chua^{1,2}, Jana Zaumseil¹, Jui-Fen Chang¹, Eric C.-W. Ou³, Peter K.-H. Ho^{1,2}, Henning Sirringhaus¹ & Richard H. Friend¹

¹Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK

²Department of Physics, National University of Singapore, Lower Kent Ridge Road, S117542, Singapore

³Institute of Materials Research and Engineering, 3 Research Link, S117602, Singapore

Organic semiconductors have been the subject of active research for over a decade now, with applications emerging in light-emitting displays and printable electronic circuits. One characteristic feature of these materials is the strong trapping of electrons but not holes¹: organic field-effect transistors (FETs) typically show p-type, but not n-type, conduction even with the appropriate low-work-function electrodes, except for a few special high-electron-affinity^{2–4} or low-bandgap⁵ organic semiconductors. Here we demonstrate that the use of an appropriate hydroxyl-free gate dielectric—such as a divinyl-tetramethylsiloxane-bis(benzocyclobutene) derivative (BCB; ref. 6)—can yield n-channel FET conduction in most conjugated polymers. The FET electron mobilities thus obtained reveal that electrons are considerably more mobile in these materials than previously thought. Electron mobilities of the order of 10^{-3} to $10^{-2} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ have been measured in a number of polyfluorene copolymers and in a dialkyl-substituted poly(*p*-phenylene-vinylene), all in the unaligned state. We further show that the reason why n-type behaviour has previously been so elusive is the trapping of electrons at the semiconductor–dielectric interface by hydroxyl groups, present in the form of silanols in the case of the commonly used SiO_2 dielectric. These findings should therefore open up new opportunities for organic complementary metal-oxide semiconductor (CMOS) circuits, in which both p-type and n-type behaviours are harnessed.

Much of the research in organic FETs has traditionally focused on the semiconductor and its contacts. Despite the importance of the gate dielectric, there have been rather few reports examining practical gate dielectric systems⁷ or the dielectric–semiconductor interface itself⁸. Recently we described a robust crosslinkable BCB that can provide a high-quality hydroxyl-free interface to the organic semiconductor⁶. BCB shows a high dielectric breakdown strength exceeding 3 MV cm^{-1} , and can be solution-cast to form the ultrathin films needed for practical low-gate-voltage plastic transistors⁶. In subsequent work, we discovered unexpectedly that some of these devices exhibit ambipolar (that is, both p- and n-type) behaviour. This greatly extends the range of suitable materials for organic CMOS circuits. Many of the special high-electron-affinity (EA) and narrow-gap organic semiconductors that enabled earlier n-channel FETs turned out unfortunately to be rather susceptible to quasi-irreversible doping processes (see, for example, a theoretical discussion in ref. 9).

We used a bottom-gate FET device configuration, in which the crosslinked BCB provides the buffer gate dielectric interface to the semiconducting polymer. The schematic structure is shown in Fig. 1a. The BCB layer is coated over $\text{p}^{++}\text{-Si/SiO}_2$ substrates used here as convenient bottom-gate substrates. We used low-work-function Ca electrodes to enable determination of the true electron mobilities without having to correct for contact resistance effects. However, we have also obtained n-FETs with

higher-work-function electrodes, including Al and Au, albeit with lower apparent mobilities. Transfer and output characteristics of the devices were measured in nitrogen by a semiconductor parameter analyser. Typical characteristics of the n-FETs based on F8BT (Fig. 1b, c; for chemical structures and names, see Table 1) or on OC₁C₁₀-PPV (Fig. 1d) as semiconductor are shown here. These devices exhibit clean n-type field-effect operation with low threshold voltage, $V_T < 10$ V, and high on-off ratio, $> 10^5$. Reducing the channel length and dielectric thickness will markedly reduce the operation voltage. The inset of Fig. 1b demonstrates clean linear (quadratic) dependence of the transistor current on the gate voltage in the linear (saturation) operating regime. The electron FET mobility ($\mu_{e,\text{FET}}$) was evaluated in the saturation regime (Table 1). Linear regime mobility does not differ by more than 30% from the saturation regime mobility.

Using this device configuration, we measured electron field-effect mobilities for a broad range of polymer semiconductors. Poly(fluorene)-based (F8-based) polymers typically give $\mu_{e,\text{FET}}$ in the range of 10^{-3} to $10^{-2} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ ($\mu_{e,\text{FET}}$ in $\text{cm}^2 \text{ V}^{-1} \text{ s}^{-1}$: F8, 1×10^{-2} ; F8BT, 5×10^{-3} ; F8T2, 6×10^{-3}), whereas the poly(*p*-phenylenevinylene)-based (PPV-based) polymers tend to show somewhat lower values ($\mu_{e,\text{FET}}$ in $\text{cm}^2 \text{ V}^{-1} \text{ s}^{-1}$: PPV, 1×10^{-4} ; OC₁C₁₀-PPV, 2×10^{-3} ; MEH-PPV, 3×10^{-5} ; CN-PPV, 3×10^{-5}). Previously it has not been generally possible to obtain meaningful μ_e values for most semiconducting polymers in transistor devices or time-of-flight (TOF) measurements. This gave rise to the widely held notion that electrons are far less mobile than holes by a factor of at least 10–100. The μ_e results obtained here directly refute this notion, and decisively show that electrons are considerably more mobile than previously thought. The results presented here are for unaligned films. With proper chain alignment, as through the liquid-crystalline phase¹⁰, we expect even higher electron mobilities to be reached.

Using Au source-drain contacts instead of Ca, we can measure the corresponding hole field-effect mobility ($\mu_{h,\text{FET}}$) in many cases and establish for these polymers that electrons are at least as mobile (if not more mobile) than holes (see Table 1). We were not able to obtain hole mobilities on F8BT and CN-PPV devices, perhaps because of a large hole injection barrier or severe hole trapping.

We further established using Fourier-transform infrared (FTIR) spectrometry that the concentration of the α -carbonyl defects in the PPV family does not exceed $5 \times 10^{11} \text{ cm}^{-2}$ at the interface. Therefore the concentration of these known electron traps is more than one order of magnitude below the field-induced charge carrier concentration (10^{13} cm^{-2}), and the μ_e values obtained are not compromised. For the PPVs at least, the existence of highly mobile electrons is consistent with previous deductions from electroluminescence radiation pattern modelling of efficient light-emitting diodes¹¹; it is also consistent with measurements of space-charge-limited currents in MEH-PPV¹² that yielded comparable room-temperature μ_e and μ_h values.

So it is now firmly established that electrons are mobile in many conjugated polymers, and that n-FETs can quite readily be realized (with the possible exception of triarylamine copolymers for which we have not been able to obtain n-FETs). This reconciles electron transport at interfaces with that in the bulk, the latter governing the efficiency of polymer light-emitting diodes¹³ in which electrons need to traverse the thickness of the light-emitting film to recombine with holes away from the metal cathode where a number of critical quenching processes operate.

n-FET conduction is not limited only to BCB as the gate dielectric. We have obtained n-FETs from other dielectrics including polyethylene (see below), poly(methyl methacrylate) and parylene (see Supplementary Fig. 1), although BCB is a preferred choice because of the complete absence of hydroxyl groups, and the other desirable attributes outlined earlier. The question then arises as to

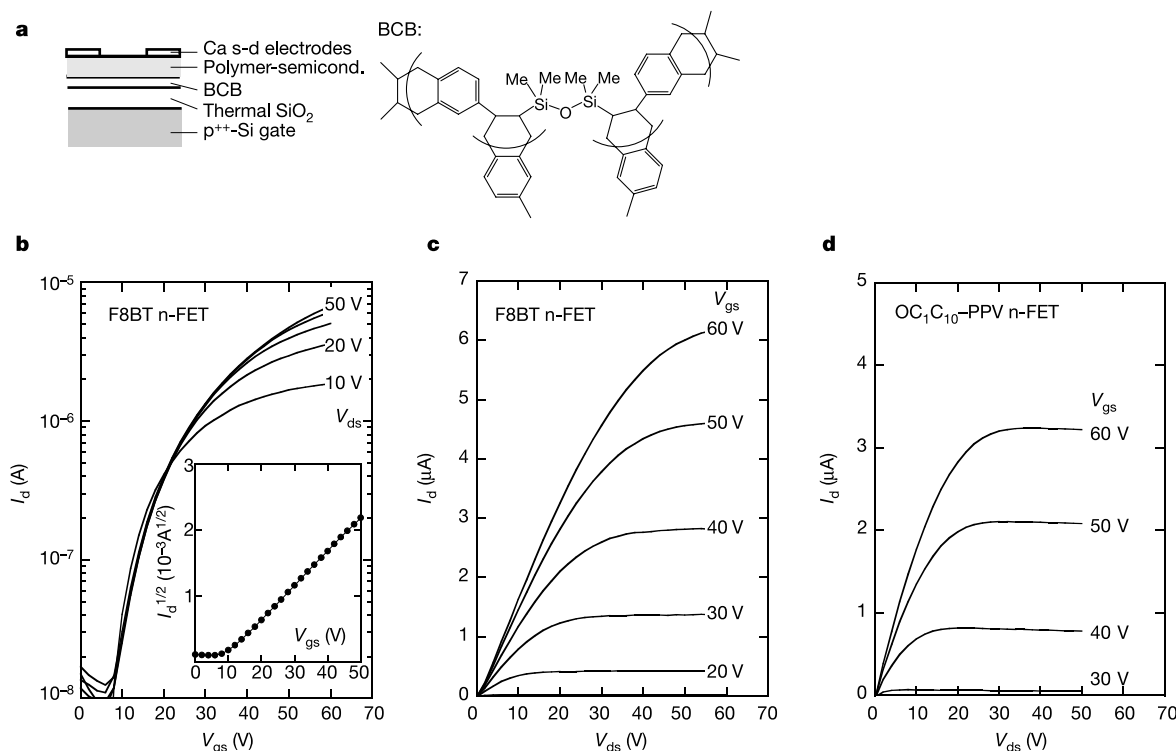
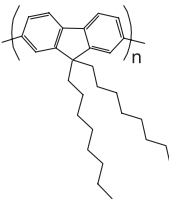
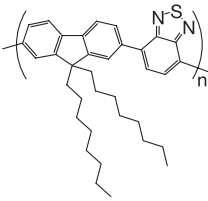
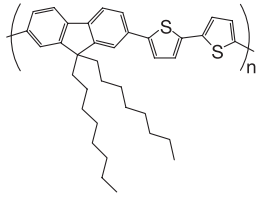
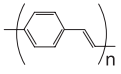
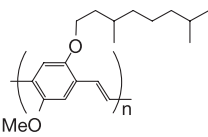
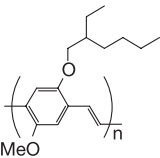
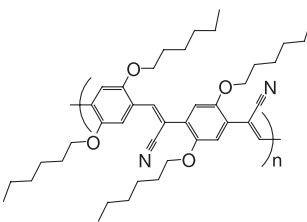
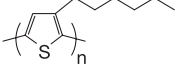


Figure 1 F8BT and OC₁C₁₀-PPV n-channel FETs with BCB/SiO₂ dielectric and Ca source-drain electrodes. **a**, Diagram of the n-FET and the chemical structure of the crosslinked BCB dielectric. s-d, source-drain. **b**, Transfer characteristics of an F8BT n-FET: channel length $L = 25 \mu\text{m}$, channel width $w = 2.5 \text{ mm}$ and gate capacitance $C_1 = 13 \text{ nF cm}^{-2}$ (SiO₂ thickness, 200 nm; BCB thickness, 50 nm). Inset shows a

well-behaved linear plot of $I_d^{1/2}$ versus V_{gs} . Values of $\mu_{e,\text{FET}}$ extracted from the linear and the saturation regimes are $7 \times 10^{-3} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and $5 \times 10^{-3} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, respectively. **c**, Output characteristics show well-defined saturation behaviour. **d**, Output characteristics of an OC₁C₁₀-PPV n-FET: $L = 200 \mu\text{m}$, $w = 10 \text{ cm}$ and $C_1 = 9 \text{ nF cm}^{-2}$.

Table 1 Mobilities for unaligned semiconducting polymers

Semiconducting polymer	Physical properties*	$\mu_{h,FET}$ (cm ² V ⁻¹ s ⁻¹)†	$\mu_{e,FET}$ (cm ² V ⁻¹ s ⁻¹)†
F8 Poly(9,9-dioctylfluorene) 	LUMO 2.4 eV HOMO 5.7 eV T_g 100 °C T_k 170 °C T_m 280 °C	3×10^{-4} (mixed xylenes, 200 °C, HMDS-SiO ₂ , Au) ²⁵	5×10^{-3} (mixed xylenes, 130 °C, BCB, Ca) 1×10^{-2} (mixed xylenes, 240 °C, BCB, Ca)
F8BT Poly(9,9-dioctylfluorene- <i>alt</i> -benzothiadiazole) 	LUMO 3.3 eV HOMO 5.9 eV T_g 130 °C T_k 220 °C T_m 250 °C	NA	4×10^{-3} (mixed xylenes, 130 °C, BCB, Ca) 5×10^{-3} (mixed xylenes; 220 °C, BCB, Ca)
F8T2 Poly(9,9-dioctylfluorene- <i>alt</i> -bithiophene) 	LUMO 3.1 eV HOMO 5.5 eV T_g 110 °C T_k 265 °C T_m 300 °C	5×10^{-3} (mixed xylenes, 130 °C, HMDS-SiO ₂ , Au) ¹⁰	6×10^{-3} (mixed xylenes, 130 °C, BCB, Ca)
PPV Poly(<i>p</i> -phenylenevinylene) 	LUMO 2.7 eV HOMO 5.2 eV	NA	1×10^{-4} (methanol, 250 °C, BCB, Ca)
OC₇C₁₀-PPV Poly(2-methoxy-5-(3,7-dimethyloctoxy)- <i>p</i> -phenylene vinylene) 	LUMO 2.8 eV HOMO 5.0 eV	5×10^{-4} (mixed xylenes, 200 °C, BCB, Au)	8×10^{-5} (1:4 tetrahydro-furan:mixed xylenes, 130 °C, BCB, Ca) 2×10^{-3} (mixed xylenes, 200 °C, BCB, Ca)
MEH-PPV Poly(2-methoxy-5-(2-ethylhexoxy)-1,4-phenylene vinylene) 	LUMO 2.8 eV HOMO 5.0 eV	5×10^{-5} (mixed xylenes, 200 °C, BCB, Au)	3×10^{-5} (mixed xylenes, 200 °C, BCB, Ca)
CN-PPV Poly(2,5-dihexoxy- α,α' -dicyano- <i>p</i> -xylylidene- <i>alt</i> -2,5-dihexoxy- <i>p</i> -xylylidene) 	LUMO 3.2 eV HOMO 5.4 eV	NA	4×10^{-5} (mixed xylenes, 130 °C, BCB, Ca)
P3HT Regioregular poly(3-hexylthiophene) 	LUMO 2.7 eV HOMO 4.9 eV T_m 235 °C	2×10^{-4} (mixed xylenes, 100 °C, BCB, Ca)	6×10^{-4} (mixed xylenes, 100 °C, BCB, Ca)

Values are electron and hole field-effect transistor mobilities for unaligned semiconducting polymers, measured in this work unless otherwise noted.
Highest-occupied-molecular-orbital (HOMO) energy values were deduced from cyclic voltammetry, and lowest-unoccupied-molecular-orbital (LUMO) energy values were estimated from the π - π^ gap, together with HOMO values.
†FET saturation regime mobility values are given for: semiconductor deposition solvent, semiconductor anneal temperature, dielectric material and source-drain electrode. BCB, divinyltetramethylsiloxane-bis(benzocyclobutene); HMDS-SiO₂, hexamethyldisilazane-treated SiO₂. T_g , glass transition temperature; T_k , crystal to liquid-crystal transition temperature; T_m , isotropic melting temperature; NA, not available. 'Mixed xylenes' is a commercial mixture of xylene isomers (predominantly the *p*-isomer) and ethylbenzene.

why n-FETs have been so elusive in the past, when primarily thermal SiO₂ or hydroxyl-containing polymers, such as poly(vinyl phenol) and polyimides (due to residual COOH groups), have been used as gate dielectrics.

Using multiple-reflection attenuated-total-reflection FTIR spectrometry (ATR-FTIR), we have found evidence for electrochemical trapping of electrons by SiOH silanol groups. The surface SiOH concentration on standard dry thermal SiO₂ grown at 800–1,000 °C has been established to be $\sim(3\text{--}7) \times 10^{13} \text{ cm}^{-2}$ (about 6–12% equivalent of a chemisorbed monolayer) in a number of studies^{14,15}. This minute concentration is nevertheless electronically significant, as it greatly exceeds the typical carrier concentration of 10^{13} cm^{-2} found in normal FET operation. We used a trapezoidal Si prism as both the gate electrode and the ATR element (Fig. 2a) to obtain the desired submonolayer interface sensitivity. A 240-nm-thick gate oxide was thermally grown into the prism, and the interdigitated Au-on-Cr source-drain electrode array was then evaporated through a shadow mask with long channel length (250 μm) to avoid scattering of the infrared radiation. A polyfluorene semiconducting polymer film was then spin-cast, and the FET device equilibrated overnight starting at time $t = -16 \text{ h}$ in the spectrometer under N₂ purge.

In the spectrum taken at $t = +16 \text{ h}$ just before device operation, we find a clear signature of SiO–H (stretching (ν), $3,630 \text{ cm}^{-1}$)¹⁵ present as lightly hydrogen-bonded species at submonolayer coverage. This mode mixes with SiO–H bending (δ) to give a $\nu + \delta$ combination band near $3,800\text{--}4,700 \text{ cm}^{-1}$ that is sensitive to hydrogen bonding^{16,17}. This band allows us to track silanol changes with time and device operation, when the fundamental region itself slowly becomes obscured by a drift of the interface reflectivity. Throughout the open-circuit rest period and for negative gate-biasing (that is, p-channel), curves (a)–(c) in Fig. 2b and c, we found only a small drift in the SiO–H $\nu + \delta$ region. However, as soon as a positive gate-bias is applied (that is, n-channel), the overall intensity of the SiO–H $\nu + \delta$ band decreases, and some intensity shifts from hydrogen-bonded ($4,340 \text{ cm}^{-1}$)¹⁶ to the nearly isolated species ($4,500 \text{ cm}^{-1}$). This is confirmed by a parallel loss of intensity in the SiO–H $\nu + 2\delta$ band region and an intensity shift from $5,150 \text{ cm}^{-1}$ to $5,320 \text{ cm}^{-1}$. The timescale for these changes is

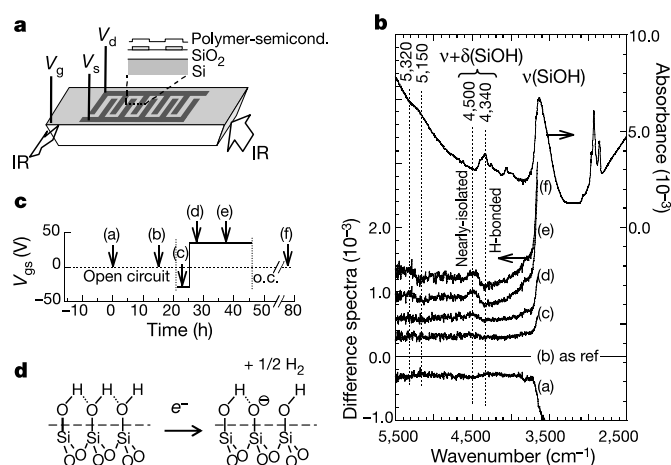


Figure 2 Evidence from vibrational spectroscopy for interfacial electron trapping on standard silicon oxide/silicon backgate device structures during attempted n-channel FET operation. **a**, Schematic structure of the FET with interdigitated Au source-drain arrays on SiO₂/Si. The device is interrogated by ATR-FTIR under nitrogen at room temperature. Nitrogen purge was started at $t = -16 \text{ h}$. **b**, The spectrum taken at time $t = +16 \text{ h}$ is plotted against the right axis. Difference ATR-FTIR spectra are plotted against the left axis, and offset for clarity for the various gate-bias conditions shown in **c**. **d**, Schematic diagram of the interfacial charge trapping mechanism in which the injected electrons are electrochemically trapped as immobile surface SiO⁻ ions.

many orders of magnitude longer than charge-carrier trapping and the associated loss of FET activity. We suspect that this is because the carriers are deeply trapped to generate SiO⁻ reaction products, which are initially confined to a narrow strip just beyond the injecting electrode, and this electrochemical reaction front then takes a long time to drift through the channel length (see Supplementary Discussion).

These changes provide direct evidence for the electrochemical OH electron-trapping mechanism shown in Fig. 2d. As a result of SiOH reduction in the n-channel regime, the interface becomes charged by a layer of immobile negative ions that compensates the gate field and pushes the gate threshold outside the measurement window. This mechanism has previously been discussed in connection with electron-trapping in hydrated thermal SiO₂ (ref. 18) and with threshold voltage offsets in silicon MOS structures¹⁹. SiOH groups are in fact generally rather acidic (acid dissociation constant, $K_a \approx 10^{-5}\text{--}10^{-4}$). The data presented here provide direct evidence that silanol groups present at the SiO₂ dielectric interface can indeed quench the n-channel FET activity of organic semiconductors that do not have sufficiently large EAs. An examination of the literature suggests an EA threshold lies at about 3.85 eV: materials with EAs only larger than this have shown n-FET behaviour on pristine SiO₂ interfaces, albeit with generally very high threshold voltages²⁰.

Using alkyl self-assembled monolayers (SAMs) on SiO₂, we have obtained further evidence in support of this interfacial trap mechanism (Fig. 3). FETs were fabricated on SiO₂ gate dielectric that was surface-passivated with alkyl-SAMs of various lengths²¹: hexamethyldisilazane (C₁), decyltrichlorosilane (C₁₀) and octadecyltrichlorosilane (C₁₈). The transfer characteristics for the second gate voltage (V_{gs}) sweep are shown after an initial sweep to 'fill' some of the traps. Unpassivated SiO₂ shows no n-FET activity at all. For the alkyl-SAM-passivated SiO₂ surface, we obtained a short-lived n-channel field effect with the gate threshold shifting to ever higher voltages during device operation. Longer-chain SAMs

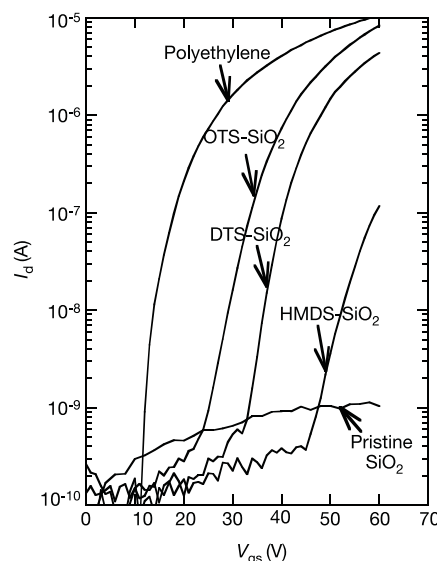


Figure 3 F8BT n-channel FETs with various siloxane self-assembled monolayers (SAMs) on SiO₂ as dielectric, or with polyethylene as buffer dielectric. Transfer characteristics with drain voltage $V_{ds} = 60 \text{ V}$ are plotted for the second gate-voltage (V_{gs}) sweep. The current below threshold is from gate leakage. For the untreated SiO₂ interface, charge trapping is so severe that no n-type behaviour can be found. With the SAM passivation (HMDS, hexamethyldisilazane; DTS, decyltrichlorosilane; OTS, octadecyltrichlorosilane), n-channel activity can be observed. However, the devices still exhibit a considerable V_{gs} threshold shift with sweep number, which is particularly severe for the shorter SAMs. For polyethylene as buffer dielectric, stable n-FET conduction is observed, as for BCB as dielectric.

suppress the rate of this shift. However, traps continue to fill and transistor current eventually disappears (see Supplementary Fig. 2). These observations indicate that traps originate at or near the SiO₂ interface. These traps cannot be completely passivated because siloxane-terminated SAMs cannot completely eliminate surface SiOH groups²². The SAM layer however provides a tunnel barrier²³ to such sites. When polyethylene was used as the buffer dielectric, stable n-channel conduction was observed, as is the case with BCB.

A further line of evidence in support of negative charge centres being generated at the SiO₂ interface upon electron injection comes from studies of gate-threshold shift in p-channel FETs with unpassivated SiO₂ as the dielectric (see Supplementary Fig. 3). When these devices are positive-gate-biased, the onset for p-channel conduction shifts to positive gate voltages. This effect is irreversible, which excludes the movement of anion impurities as its cause. It is consistent however with the formation of interface SiO⁻ according to our mechanism, which then shifts the p-channel gate threshold to positive voltages.

Finally, for some polymers and device configurations with BCB dielectric we have obtained ambipolar behaviour. Figure 4a shows the transfer characteristics of a top-gated F8T2 FET with Au (on a Cr adhesion layer) source-drain electrodes. Ambipolar operation is achieved despite the sizeable mismatch estimated by the work-function of 'clean Au' (5.1 eV) and the lowest-unoccupied molecular orbital (LUMO) of F8T2 (3.1 eV), which anticipates a large electron injection barrier ($\Delta_e = 2.0$ eV) in a zero-order vacuum alignment picture. The fact that electron injection can still be obtained suggests that the barrier for electron injection is considerably lowered, owing perhaps to the formation of interface dipoles. Nevertheless Au is still a relatively poor electron injector, which explains the lower current during n-channel operation. Figure 4b shows an ambipolar bottom-gated regioirregular-P3HT FET with Ca source-drain electrodes. Both electrons and holes can be injected through this contact, as the π - π^* gap of P3HT is relatively small (1.9 eV). The $\mu_{e,FET}$ and $\mu_{h,FET}$ values are similar (10^{-4} to 10^{-3} cm² V⁻¹ s⁻¹), and the $\mu_{h,FET}$ value is in fact similar also to the one obtained with Au electrodes at this interface (4×10^{-4} cm² V⁻¹ s⁻¹), indicating that Ca does not appear to

impose a large hole injection barrier. However, these values are still considerably lower than the best hole-mobility value obtained for this type of material on SiO₂ dielectric²⁴, which suggests that there may be scope for further optimization of the P3HT chain packing at this interface to produce even higher ambipolar mobilities.

We have demonstrated the ubiquity of n-channel organic FET operation with a wide range of polymers previously thought to show only p-channel activity. Our findings should also be relevant to other cases where inorganic oxide surfaces are put in contact with molecular semiconductors—for example, in photovoltaic applications, and in single-molecule devices. □

Methods

Bottom-gated polymer FETs with BCB as buffer gate dielectric

Thermal p⁺⁺-Si/SiO₂ (200-nm-thick) substrates were spin-coated with a 50–80-nm-thick film of the BCB monomer from a mesitylene solution. (The BCB sample was first purified to extract out the 1,2-dihydro-2,2,4-trimethylquinoline stabilizer, which can act as traps.) After crosslinking to the final gate dielectric by rapid thermal anneal on a hotplate at 290 °C (15 s, in a nitrogen-filled glove-box), a 70-nm-thick semiconducting polymer film was deposited by spin-coating in the glove-box. The polymers were thus deposited in the unaligned state and briefly annealed at the stated temperature in the glove-box. Root-mean-square roughness of this dielectric interface is 3 Å (2 × 2 μm). To form ohmic source and drain contacts, 100-nm-thick Ca electrodes were deposited through a shadow mask to provide a channel length of 25–200 μm and a channel width of 2.5–10 mm. To protect Ca from oxidation, the devices were then encapsulated with a 30-nm silicon monoxide layer.

Bottom-gated polymer FETs with polyethylene as buffer gate dielectric

As above, but medium-density polyethylene in decalin was spun at 190 °C onto the substrates and annealed at 90 °C for 3 h. Root-mean-square roughness of the dielectric interface produced in this way is 24 Å (2 × 2 μm).

Surface passivation of SiO₂ by self-assembled-monolayers

Thermal p⁺⁺-Si/SiO₂ (200-nm-thick) substrates were treated with hexamethyldisilazane vapour at 120 °C for 5 h and then annealed in air at 120 °C for 2 min. For the C₁₀ and C₁₈ alkyl-chain SAMs, the substrates were treated with 5 mM decyltrichlorosilane or octadecyltrichlorosilane in hexane for 24 h, washed with hexane and then annealed in air at 120 °C for 2 min.

FTIR spectrometric analysis of carbonyl (C = O) concentration

10-μm-thick PPV, OC₁₀-PPV and CN-PPV films were cast onto intrinsic Si wafers and their FTIR spectra collected in transmission. The integrated intensity of the C=O stretch at ~1,696 cm⁻¹ is compared to those of the in-plane phenylene ring vibrations at ~1,510 and 1,680 cm⁻¹. After calibrating with the empirical infrared cross-sections, we obtained the upper limit for the C=O concentration in these polymers to be (in cm⁻³): PPV, <4 × 10¹⁸; OC₁₀-PPV, <1 × 10¹⁷; and CN-PPV, <4 × 10¹⁷. The field-induced charge-carriers reside within 2 nm of the interface. Assuming uniform distribution, we computed the interface carbonyl density in this interfacial layer: PPV, <8 × 10¹¹; OC₁₀-PPV, <2 × 10¹⁰; and CN-PPV, <8 × 10¹⁰ cm⁻². For comparison, the interface charge density at V_{gs} of 100 V is 8 × 10¹² cm⁻², which is at least 1–2 orders of magnitude larger than the C=O density, and thus C=O is not expected to play a role in limiting the reported $\mu_{e,FET}$ values. Because combination overtone bands also reside in this C=O region, our estimates here are conservative.

Received 21 July 2004; accepted 14 January 2005; doi:10.1038/nature03376.

1. Dimitrakopoulos, C. D. & Malenfant, P. R. L. Organic thin film transistors for large area electronics. *Adv. Mater.* **14**, 99–117 (2002).
2. Dodabalapur, A., Katz, H. E., Torsi, L. & Haddon, R. C. Organic heterostructure field-effect transistors. *Science* **269**, 1560–1562 (1995).
3. Katz, H. E. *et al.* A soluble and air-stable organic semiconductor with high electron mobility. *Nature* **404**, 478–481 (2000).
4. Babel, A. & Jenekhe, S. A. High electron mobility in ladder polymer field-effect transistors. *J. Am. Chem. Soc.* **125**, 13656–13657 (2003).
5. Meijer, E. J. *et al.* Solution-processed ambipolar organic field-effect transistors and inverters. *Nature Mater.* **2**, 678–682 (2003).
6. Chua, L. L., Ho, P. K. H., Sirringhaus, H. & Friend, R. H. High stability ultrathin spin-on benzocyclobutene gate dielectric for polymer field-effect transistors. *Appl. Phys. Lett.* **84**, 3400–3402 (2004).
7. Veres, J., Ogier, S. D., Leeming, S. W., Cupertino, D. C. & Khaffaf, S. M. Low-k insulators as the choice of dielectrics in organic field-effect transistors. *Adv. Funct. Mater.* **13**, 199–204 (2003).
8. Chua, L. L., Ho, P. K. H., Sirringhaus, H. & Friend, R. H. Observation of field-effect behaviour at self-organised interfaces. *Adv. Mater.* **16**, 1609–1615 (2004).
9. Brown, A. R., De Leeuw, D. M., Lous, E. J. & Havinga, E. E. Organic n-type field-effect transistors. *Synth. Met.* **66**, 257–261 (1994).
10. Sirringhaus, H. *et al.* Mobility enhancement in conjugated polymer field-effect transistors through chain alignment in a liquid-crystalline phase. *Appl. Phys. Lett.* **77**, 406–408 (2000).
11. Kim, J. S., Ho, P. K. H., Greenham, N. C. & Friend, R. H. Electroluminescence emission pattern of organic light-emitting diodes: Implications for device efficiency calculations. *J. Appl. Phys.* **88**, 1073–1081 (2000).

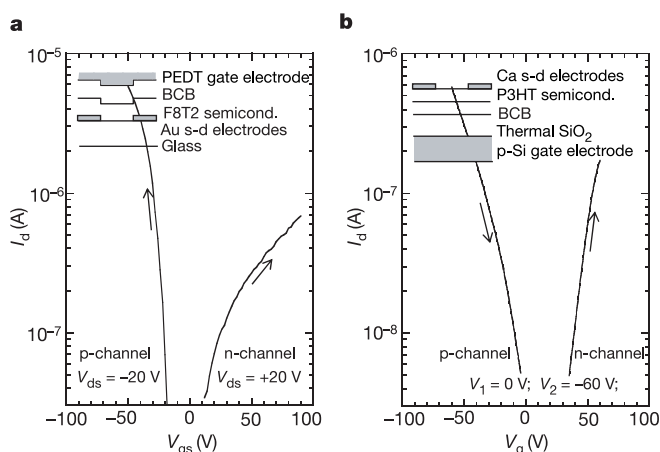


Figure 4 Ambipolar F8T2 and regioirregular-P3HT FETs. **a**, Top-gated F8T2 ambipolar-FET with Au (on Cr adhesion layer) source-drain electrodes and PEDT:PSSR as top gate. $L = 3$ μm; $w = 10$ mm; BCB thickness, 200 nm; $C_i = 10$ nF cm⁻². The data were collected in two parts with $V_{ds} = -20$ V for p-channel scan; $V_{ds} = +20$ V for n-channel scan. Electron and hole operation is achieved despite the sizeable mismatch between the work-function of 'clean Au' (5.1 eV) and the LUMO of F8T2 (3.1 eV). **b**, Bottom-gated regioirregular-P3HT ambipolar-FET with Ca source-drain electrodes and Si as bottom gate. $L = 200$ μm; $w = 10$ cm; BCB thickness, 70 nm; $C_i = 9$ nF cm⁻². The data were collected in one sweep (as the second half of a sweep cycle) with one of the source-drain electrodes at $V_1 = 0$ V, and the other one at $V_2 = -60$ V.

12. Bozano, L., Carter, S., Scott, J., Malliaras, G. & Brock, P. Temperature- and field-dependent electron and hole mobilities in polymer light-emitting diodes. *Appl. Phys. Lett.* **74**, 1132–1134 (1999).
13. Ho, P. K. H. *et al.* Molecular-scale interface engineering for polymer light-emitting diodes. *Nature* **404**, 481–484 (2000).
14. Nagasawa, Y. *et al.* The study of the thermal oxide films on silicon wafers by Fourier transform infrared attenuated total reflection spectroscopy. *J. Appl. Phys.* **68**, 1429–1434 (1990).
15. Dugas, V. & Chevalier, Y. Surface hydroxylation and silane grafting on fumed and thermal silica. *J. Colloid Interf. Sci.* **264**, 354–361 (2003).
16. Perry, C. C. & Li, X. Structural studies of gel phases. *J. Chem. Soc. Faraday Trans.* **87**, 3857–3862 (1991).
17. Tardif, F., Chabli, A., Kanel, A., Rochat, N. & Veillerot, M. Thermal evolution of chemical oxides and (100) silicon at 300 °C in ambient air as seen by attenuated total reflection infrared spectroscopy. *J. Electrochem. Soc.* **150**, G333–G338 (2003).
18. Nicollan, E. H., Berglund, C. N., Schmidt, P. F. & Andrews, J. M. Electrochemical charging of thermal SiO₂ films by injected electron currents. *J. Appl. Phys.* **42**, 5654–5664 (1971).
19. Hartstein, A. & Young, D. R. Identification of electron traps in thermal silicon dioxide films. *Appl. Phys. Lett.* **38**, 631–633 (1981).
20. Malenfant, P. R. L. *et al.* N-type organic thin-film transistor with high field-effect mobility based on a N,N'-dialkyl-3,4,9,10-perylene tetracarboxylic diimide derivative. *Appl. Phys. Lett.* **80**, 2517–2519 (2002).
21. Salleo, A., Chabiny, M. L., Yang, M. S. & Street, R. A. Polymer thin-film transistors with chemically modified dielectric interfaces. *Appl. Phys. Lett.* **81**, 4383–4385 (2002).
22. Angst, D. L. & Simmons, G. W. Moisture absorption characteristics of organosiloxane self-assembled monolayers. *Langmuir* **7**, 2236–2242 (1991).
23. Wasserman, S. R., Tao, Y. T. & Whitesides, G. M. Structure and reactivity of alkylsiloxane monolayers formed by reaction of alkyltrichlorosilanes on silicon substrates. *Langmuir* **5**, 1074–1087 (1989).
24. Sirringhaus, H. *et al.* Two-dimensional charge transport in self-organised, high-mobility conjugated polymers. *Nature* **401**, 685–688 (1999).
25. Wilson, R. J. *Polymer Field-effect Transistors from Polyfluorene-based Conjugated Polymers*. PhD thesis, Univ. Cambridge (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Achen for providing the BCB material, which is commercially available as Dow Cyclotene, and Merck KGaA for P3HT. L.-L.C. and J.-F.C. thank the EPSRC Carbon-Based Electronics programme for support, and J.Z. thanks the Gates Cambridge Trust for support. P.K.-H.H. thanks St John's College for a research fellowship, the National University of Singapore for research support, and S. J. Chua for technical discussions, support and access to laboratory facilities. We also thank C.-C. Chum for technical support.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com.

Correspondence and requests for materials should be addressed to P.K.-H.H. (phyhop@nus.edu.sg) and R.H.F. (rhf10@cam.ac.uk).

Ultrafast memory loss and energy redistribution in the hydrogen bond network of liquid H₂O

M. L. Cowan^{1*}, B. D. Bruner^{1*}, N. Huse^{2*}, J. R. Dwyer¹, B. Chugh¹, E. T. J. Nibbering², T. Elsaesser² & R. J. D. Miller¹

¹Departments of Chemistry and Physics, University of Toronto, 80 St George Street, Toronto, Ontario, Canada M5S3H6

²Max-Born-Institut für Nichtlineare Optik und Kurzzeitspektroskopie, Max-Born-Strasse 2A, D-12489 Berlin, Germany

* These authors contributed equally to this work

Many of the unusual properties of liquid water are attributed to its unique structure, comprised of a random and fluctuating three-dimensional network of hydrogen bonds that link the highly polar water molecules^{1,2}. One of the most direct probes of the dynamics of this network is the infrared spectrum of the OH stretching vibration^{3–11}, which reflects the distribution of hydrogen-bonded structures and the intermolecular forces controlling the structural dynamics of the liquid. Indeed, water dynamics has been studied in detail^{12–14}, most recently using multi-dimensional nonlinear infrared spectroscopy^{15,16} for acquiring structural and dynamical information on femtosecond timescales. But owing to technical difficulties, only OH

stretching vibrations in D₂O or OD vibrations in H₂O could be monitored. Here we show that using a specially designed, ultra-thin sample cell allows us to observe OH stretching vibrations in H₂O. Under these fully resonant conditions, we observe hydrogen bond network dynamics more than one order of magnitude faster than seen in earlier studies that include an extremely fast sweep in the OH frequencies on a 50-fs timescale and an equally fast disappearance of the initial inhomogeneous distribution of sites. Our results highlight the efficiency of energy redistribution within the hydrogen-bonded network, and that liquid water essentially loses the memory of persistent correlations in its structure within 50 fs.

The high optical density of the OH stretching mode in pure H₂O and parasitic window signals in conventional samples have so far

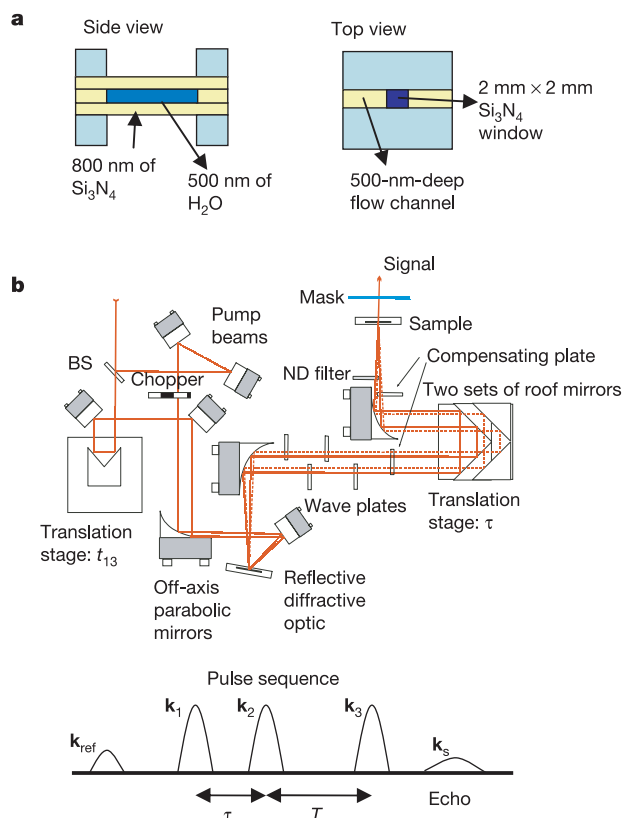


Figure 1 Experimental set-up. **a**, Nanofluidic sample cell: Each Si₃N₄ window has a 2 mm × 2 mm clear aperture and is 800 nm thick, which provides the necessary structural support for a stable sample, yet is thin enough to give a negligible non-resonant window signal. The windows were also coated with a 10-nm SiO₂ layer to ensure that their surfaces are sufficiently hydrophilic to allow effective capillary flow of the water into the cell and uniform distribution of the water once inside. **b**, Schematic of two-dimensional infrared experiment: The 70-fs pulse from a mid-IR Optical Parametric Amplifier is split into two at a beam splitter (BS), and one is given a variable delay t_{13} using a retroreflector on a translation stage. The two beams are focused onto a reflective diffractive optic at a small angle to each other, using an off-axis parabolic mirror. The diffractive optic is optimized to diffract 70% of the intensity of each beam into the ± 1 diffraction orders. The resulting four beams lie at the corners of a square, forming a 'boxcar' pattern. Using the second translation stage, the bottom two beams are then delayed by a time τ . This results in the pulse sequence shown, with the two pump beams (k_1, k_2) separated by the coherence time τ and the second pump and probe beam (k_2, k_3) separated by the population time T . The fourth beam in the boxcar pattern is the weak reference pulse k_{ref} , which is attenuated by the neutral density (ND) filter and passed through the sample cell before the excitation pulses (k_1, k_2, k_3). The 1- μ J pulses are then focused to a 150- μ m spot in the sample, and the signal is heterodyne detected, using the fourth beam as a local oscillator. Under these conditions, the temperature rise in the sample measured using a thermal imaging camera was less than 8 K above room temperature.

hindered attempts to access the fastest relaxation processes of liquid water in pure H₂O. To overcome these problems, we designed the nanofluidic cell illustrated in Fig. 1a—extremely thin (800 nm) Si₃N₄ windows remove nonlinear window signals, while the similarly thin (500 nm) water layer has a peak optical density of only 0.2. The experiment uses our spectroscopy method based on diffractive optics¹⁷ and involves applying three ~ 1 - μ J, 70-fs pulses centred at 3,350 cm⁻¹ in the OH stretching band to generate photon echoes as a function of the coherence time τ and the population time T , as shown in Fig. 1b. The photon echo signal is heterodyne-detected by spectral interferometry^{17–19} with a reference pulse from the same mid-infrared source, directly giving the echo as a function of the detected frequency dimension ν_3 . The coherence time τ is then scanned at constant population time T and the signal is Fourier-transformed along the τ dimension to produce the excitation frequency dimension ν_1 . The absorptive part of the resulting two-dimensional infrared spectrum is determined by fitting phase factors to independently measured spectrally resolved pump–probe data^{12,17,19}.

Figure 2 shows spectrally integrated transient grating measurements for population times T up to 2 ps for both parallel and perpendicular polarizations. Both polarization conditions exhibit a fast exponential decay with a 95-fs time constant, followed by a rise with a 1.3-ps time constant at later times T . On this timescale, energy transfer processes can influence the polarization anisotropy⁶ (Fig. 2), that is, energy transfer from the preferentially excited OH vibrations along the excitation polarization direction to different water molecules leads to a decay in the degree of polarization of the signal. The polarization anisotropy decays exponentially with a 75-fs time constant, substantially shorter than the 700-fs time constant we measure in a diluted 6:1 D₂O:H₂O mixture under the same conditions in our nanofluidic cell (see Supplementary Information); this 700-fs time constant is in agreement with pump–probe measurements previously taken with 200-fs pulses⁶.

By heterodyne detecting and spectrally resolving the transient grating signal, we gain much more detailed information about the dynamics of the system. Figure 3 shows the absorptive part of the spectrally resolved transient grating signal with the polarization of all three pulses parallel. At $T = 0$, the positive peak corresponds to reduced absorption (bleaching) caused by the depletion of the $\nu = 0$ state of the OH stretching oscillator and stimulated emission from the $\nu = 1$ state. The negative peak corresponds to excited state

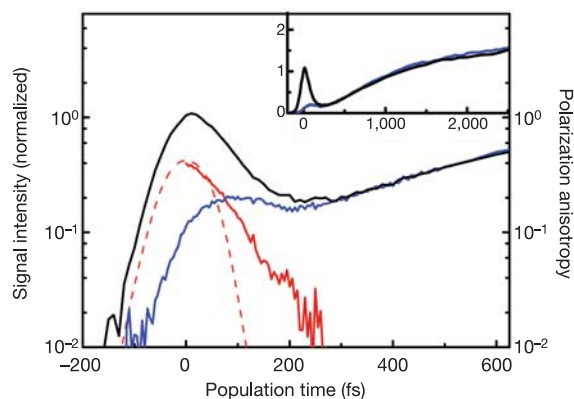


Figure 2 Spectrally integrated transient grating data in pure H₂O. The parallel polarization [0°, 0°, 0°] (black line) and crossed probe polarization [0°, 0°, 90°, 90°] (blue line) data show a fast initial decay of the population, followed by a thermal grating rise due to vibrational relaxation on a picosecond timescale. The dashed red line shows the pulse-width limited response that is convolved with a bi-exponential to determine the decay rates. The polarization anisotropy (red line), shows the fast sub-100 fs energy transfer of pure H₂O. Inset, longer time dynamics. The initial value of the polarization anisotropy is 0.4, as expected for the excitation of individual linear transition dipole moments.

absorption of the $\nu = 1$ to $\nu = 2$ transition, which is red-shifted from the $\nu = 0$ to $\nu = 1$ transition, owing to the strong anharmonicity of the OH stretching oscillator. The time evolution of such signals depends strongly on the spectral position. The enhanced absorption below 3,000 cm⁻¹ and the bleaching above 3,500 cm⁻¹ display a very fast decay that is close to the 50-fs time resolution of our experiment. A substantially slower decay and a subsequent picosecond timescale rise are observed between 3,170 and 3,400 cm⁻¹, at the centre of the bleaching contribution. Such comparably slow components dominate the spectrally integrated kinetics shown in Fig. 2.

Two-dimensional infrared spectra at different population times are shown in Fig. 4. The absorptive component of the echo signal is plotted. At $T = 0$, such spectra display an on-diagonal peak due to bleaching and stimulated emission of the $\nu = 0$ to $\nu = 1$ transition and an off-diagonal absorption peak due to the $\nu = 1$ to $\nu = 2$ transition. At $T = 0$, the on-diagonal peak is stretched along the diagonal, indicating inhomogeneous broadening. At $T = 50$ fs, this inhomogeneity is almost entirely lost (>90% based on curvature of the line contours), and by $T = 100$ fs it is completely gone. The excited-state peak off the diagonal in Fig. 4 decays on the same timescale, in agreement with the transient grating results of Fig. 3. Most importantly, there are extremely fast processes that wash out the structural variations (inhomogeneous broadening), such that the OH stretching excitation loses its memory on extraordinarily fast timescales, faster than in any other liquid.

Liquid water represents a disordered ensemble of highly polar molecules linked through a fluctuating network of intermolecular hydrogen bonds. The transition frequency of an individual OH stretching oscillator in this system depends on the local environment, resulting in an inhomogeneous broadening of the OH stretching band of the order of several hundreds of cm⁻¹. In addition, the OH stretching oscillators on different water molecules are coupled through hydrogen bonds and through other mechanisms, for example, dipole–dipole coupling, that lead to rapid

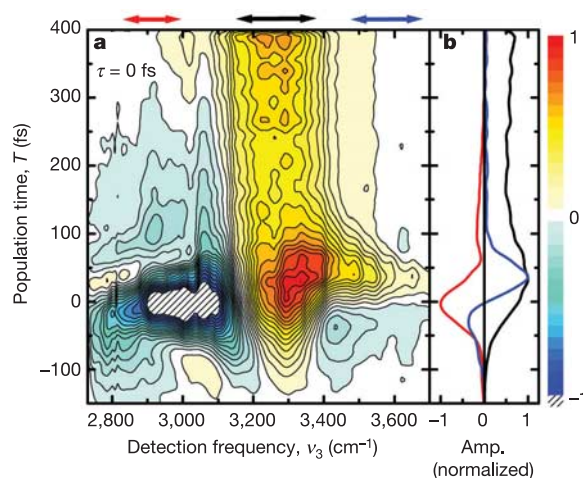


Figure 3 Absorptive component of the spectrally resolved transient grating signal, plotted as a function of population time T . **a**, This shows the fast spectral diffusion and excited state decay, and the beginning of the thermal grating rise due to vibrational relaxation. **b**, Normalized amplitude (Amp.) of the transient grating signal spectrally integrated in three windows, on the red (red line), middle (black line) and blue (blue line) sides of the spectrum. The data are not corrected for the slight frequency chirp of the femtosecond pulses. Independent pulse characterization by frequency-resolved optical gating shows that the signals rise within the time resolution of the experiment at all detection frequencies, that is, also for detection frequencies above 3,400 cm⁻¹. Spectral diffusion dominates the fast decay of the red and blue regions, while population dynamics dominate the centre of the spectrum. The contour scale is displayed from -1 to 1 in order to focus on the region of interest.

transfer of vibrational energy between neighbouring molecules. Gas-phase studies of water clusters and theoretical calculations have shown that the coupling between OH oscillators is small compared to the disorder-related inhomogeneous broadening^{20,21}. Taking this and the very short excitation pulses into consideration, we can think of the OH excitation as initially localized on a particular microscopic configuration of water molecules that then evolves over time. Nuclear motions, excitation transfer and couplings due to the highly polar nature of water open up relaxation channels and configuration-averaging mechanisms that are absent in other liquids.

The spectrally resolved transient-grating data (Fig. 3) display a very fast component on the red and blue edges, and in addition, slower kinetics in the centre of the OH stretching band. The slow component observed between 3,170 and 3,400 cm^{-1} reflect the population kinetics of the excited OH stretching oscillators. The $\tau_1 = 95$ fs decay of the transient grating signal corresponds to a depopulation of the $\nu = 1$ state with $T_1 = 2\tau_1 = 190$ fs, with possible short-lived intermediates contributing through terms connected to the ground-state recovery. The slower 1.3-ps rise of the

spectrally integrated transient-grating signal at late population times (inset of Fig. 2) is attributed to heating effects associated with vibrational relaxation, similar to the behaviour found in spectrally resolved infrared pump-probe and infrared pump-Raman-probe studies of water^{22,23}. The 75-fs decay in the polarization anisotropy (Fig. 2) is attributed to energy transfer among the OH stretching vibrations. We note that depolarization of the OH excitation requires the excitation sampling of several different waters within the pentameric structure of hydrogen-bonded water molecules and we can infer that this timescale is an upper limit for the excitation hopping time.

We observe an extremely fast sub-100-fs component of spectral diffusion below 3,000 cm^{-1} and above 3,500 cm^{-1} . Our results point to a blueshift of the $\nu = 1$ to $\nu = 2$ transition and a simultaneous red-shift of the $\nu = 0$ to $\nu = 1$ transition on a 50-fs timescale. This finding suggests ultrafast structural changes by which the dipole induced on the $\nu = 0$ to $\nu = 1$ transition is solvated²⁴, thus lowering the energy of the $\nu = 1$ state of the oscillator. This solvation process is distinct from solvent repolarization involved in electronic transitions because the induced dipole changes with OH excitation are much smaller in magnitude and there are also local interactions through hydrogen bonding. Rather, this effect is related to fluctuations local to the excited OH that change the effective degree of hydrogen bonding. The timescale of the fastest spectral diffusion processes is definitely shorter than the 200-fs period of the O-H...O hydrogen-bond mode (frequency 170 cm^{-1}) and other reorientation and energy transfer processes occurring in the subpicosecond to picosecond time domain.

We conclude that structural dynamics related to such degrees of freedom plays a minor part in the ultrafast loss of memory observed here. Instead, we consider fast librational motions (hindered rotations), which are the relevant excitations in the hydrogen-bonded network^{25–27}. The low-frequency vibrational spectrum of liquid H_2O displays a pronounced band between 300 and 1,000 cm^{-1} that originates from librational excitations with periods between 90 and 30 fs, the latter being in the relevant time range to account for the speed of both the spectral diffusion in Fig. 3 and the dephasing in Fig. 4. Energy transfer processes on a somewhat slower timescale also contribute to the observed spectral diffusion and loss of memory, although they do not dominate the process—as might be expected from extrapolations of other isotopic studies of liquid water^{6,28}.

We have studied pure H_2O to determine the microscopic processes leading to energy redistribution in the fully resonant hydrogen-bond network. This work demonstrates the loss of memory of persistent correlations in water structure within 50 fs that favours rapid relaxation of elementary excitations and is probably important in stabilizing biological systems coupled to this network of hydrogen bonds. □

Received 18 October 2004; accepted 21 January 2005; doi:10.1038/nature03383.

1. Eisenberg, D. & Kauzmann, W. *The Structure and Properties of Water* (Oxford Univ. Press, New York, 1969).
2. Franks, F. (ed.) *Water, a Comprehensive Treatise* (Plenum, New York, 1972).
3. Luzar, A. & Chandler, D. Hydrogen-bond kinetics in liquid water. *Nature* **379**, 55–57 (1996).
4. Marx, D., Tuckerman, M. E., Hutter, J. & Parrinello, M. The nature of the hydrated excess proton in water. *Nature* **397**, 601–604 (1999).
5. Graener, H., Seifert, G. & Laubereau, A. New spectroscopy of water using tunable picosecond pulses in the infrared. *Phys. Rev. Lett.* **66**, 2092–2095 (1991).
6. Woutersen, S. & Bakker, H. J. Resonant intermolecular transfer of vibrational energy in liquid water. *Nature* **402**, 507–509 (1999).
7. Gale, G. M. et al. Femtosecond dynamics of hydrogen bonds in liquid water: A real-time study. *Phys. Rev. Lett.* **82**, 1068–1071 (1999).
8. Stenger, J., Madsen, D., Hamm, P., Nibbering, E. T. J. & Elsaesser, T. Ultrafast vibrational dephasing of liquid water. *Phys. Rev. Lett.* **87**, 027401 (2001).
9. Möller, K. B., Rey, R. & Hynes, J. T. Hydrogen bond dynamics in water and ultrafast infrared spectroscopy: a theoretical study. *J. Phys. Chem. A* **108**, 1275–1289 (2004).
10. Lawrence, C. P. & Skinner, J. L. Vibrational spectroscopy of HOD in liquid D_2O . Infrared line shapes and vibrational Stokes shift. *J. Chem. Phys.* **117**, 8847–8854 (2002).
11. Torre, R., Bartolini, P. & Righini, R. Structural relaxation in supercooled water by time-resolved spectroscopy. *Nature* **428**, 296–299 (2004).

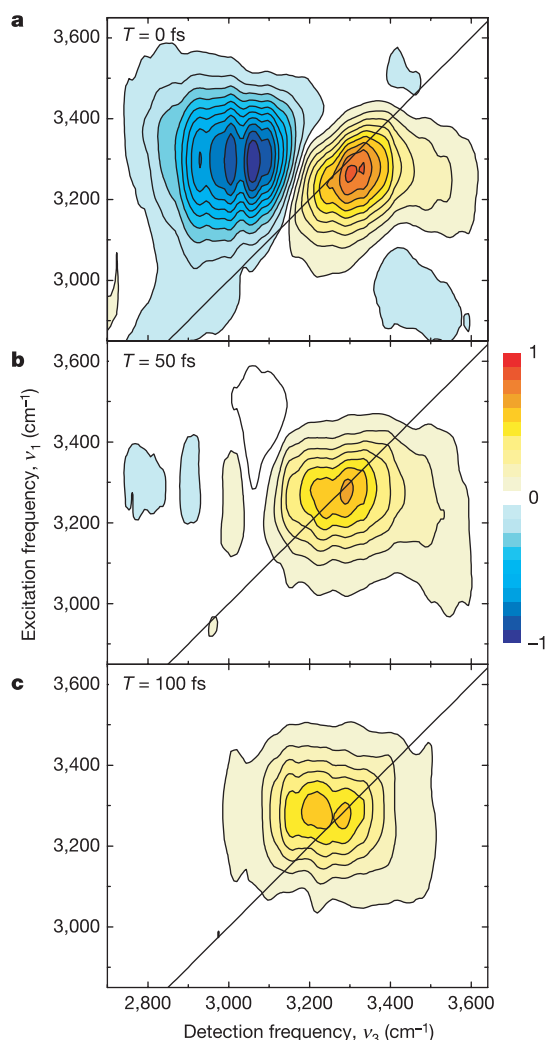


Figure 4 Absorptive components of the two-dimensional-infrared echo spectra of pure liquid H_2O for different population times. **a**, $T = 0$ fs; **b**, $T = 50$ fs; **c**, $T = 100$ fs. The inhomogeneity indicated by the stretching of the positive peak along the diagonal in the $T = 0$ spectrum has almost completely decayed by $T = 50$ fs, clearly showing the very rapid dephasing and loss of memory in the system.

12. Asbury, J. B. *et al.* Water dynamics: vibrational echo correlation spectroscopy and comparison to molecular dynamics simulations. *J. Phys. Chem. A* **108**, 1107–1119 (2004).
13. Fecko, C. J., Eaves, J. D., Loparo, J. J., Tokmakoff, A. & Geissler, P. L. Ultrafast hydrogen-bond dynamics in the infrared spectroscopy of water. *Science* **301**, 1698–1702 (2003).
14. Stenger, J., Madsen, D., Hamm, P., Nibbering, E. T. J. & Elsaesser, T. A photon echo peak shift study of liquid water. *J. Phys. Chem. A* **106**, 2341–2350 (2002).
15. Asplund, M. C., Zanni, M. T. & Hochstrasser, R. M. Two dimensional infrared spectroscopy of peptides by phase-controlled femtosecond vibrational photon echoes. *Proc. Natl Acad. Sci. USA* **97**, 8219–8224 (2000).
16. Mukamel, S. Multidimensional femtosecond correlation spectroscopies of electronic and vibrational excitations. *Annu. Rev. Phys. Chem.* **51**, 691–729 (2000).
17. Cowan, M. L., Ogilvie, J. P. & Miller, R. J. D. Two-dimensional spectroscopy using diffractive optics based phased-locked photon echoes. *Chem. Phys. Lett.* **386**, 184–189 (2004).
18. Lepetit, L., Cheriaux, G. & Joffe, M. Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy. *J. Opt. Soc. Am. B* **104**, 2467–2474 (1995).
19. Hybl, J. D., Albrecht, A. W., Faeder, S. M. G. & Jonas, D. M. Two-dimensional electronic spectroscopy. *Chem. Phys. Lett.* **297**, 307–313 (1998).
20. Rice, S. A., Bergren, M. S., Beich, A. C. & Nielson, G. A theoretical analysis of the OH stretching spectra of ice Ih, liquid water, and amorphous solid water. *J. Phys. Chem.* **87**, 4295–4308 (1983).
21. Wojcik, M. J., Buch, V. & Devlin, J. P. Spectra of isotopic ice mixtures. *J. Chem. Phys.* **99**, 2332–2344 (1993).
22. Lock, A. J. & Bakker, H. J. Temperature dependence of vibrational relaxation in liquid H₂O. *J. Chem. Phys.* **117**, 1708–1713 (2002).
23. Pakoulev, A., Wang, Z. & Dlott, D. Vibrational relaxation and spectral evolution following ultrafast OH stretch excitation of water. *Chem. Phys. Lett.* **371**, 594–600 (2003).
24. Jimenez, R., Fleming, G. R., Kumar, P. V. & Maroncelli, M. Femtosecond solvation dynamics in water. *Nature* **369**, 471–473 (1994).
25. Castner, E. W. Jr, Chang, Y. J., Chu, Y. C. & Walrafen, G. E. The intermolecular dynamics of liquid water. *J. Chem. Phys.* **102**, 653–659 (1995).
26. Saito, S. & Ohmine, I. Third order nonlinear response of liquid water. *J. Chem. Phys.* **106**, 4889–4893 (1997).
27. Pohorille, A., Pratt, L. R., LaViolette, R. A., Wilson, M. A. & MacElroy, R. D. Comparison of the structure of harmonic aqueous glasses and liquid water. *J. Chem. Phys.* **87**, 6070–6077 (1987).
28. Poulsen, J. A., Nyman, G. & Nordholm, S. Wave packet study of ultrafast relaxation in ice Ih and liquid water. Resonant intermolecular vibrational energy transfer. *J. Phys. Chem. A* **107**, 8420–8428 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. Weik for help with the use of a thermal imaging camera. Financial support by the Deutsche Forschungsgemeinschaft, the Humboldt foundation (R.J.D.M.), the Canadian Foundation for Innovation, the Natural Sciences and Engineering Research Council of Canada, and Photonics Research Ontario is acknowledged.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.J.D.M. (rjdmler@phys.chem.utoronto.ca).

The formation of cubic ice under conditions relevant to Earth's atmosphere

Benjamin J. Murray, Daniel A. Knopf & Allan K. Bertram

Department of Chemistry, University of British Columbia, 2036 Main Mall, Vancouver, British Columbia V6T 1Z1, Canada

An important mechanism for ice cloud formation in the Earth's atmosphere is homogeneous nucleation of ice in aqueous droplets, and this process is generally assumed to produce hexagonal ice^{1,2}. However, there are some reports that the metastable crystalline phase of ice, cubic ice, may form in the Earth's atmosphere^{3–5}. Here we present laboratory experiments demonstrating that cubic ice forms when micrometre-sized droplets of pure water and aqueous solutions freeze homogeneously at cooling rates approaching those found in the atmosphere. We find that the formation of cubic ice is dominant when droplets freeze at temperatures below 190 K, which is in the temperature range relevant for polar stratospheric clouds and clouds in the tropical tropopause region. These results, together with heat

transfer calculations, suggest that cubic ice will form in the Earth's atmosphere. If there were a significant fraction of cubic ice in some cold clouds this could increase their water vapour pressure, and modify their microphysics and ice particle size distributions⁵. Under specific conditions this may lead to enhanced dehydration of the tropopause region⁵.

An important mechanism for ice cloud formation in the Earth's atmosphere is homogeneous nucleation of ice in aqueous droplets. For example, polar stratospheric clouds, which play a key role in stratospheric ozone depletion, can form by homogeneous freezing of HNO₃–H₂SO₄–H₂O droplets at temperatures below 189 K (ref. 6). Also, ice cirrus clouds, which play an important role in the Earth's climate⁷, can form through the homogeneous freezing of NH₃–H₂SO₄–H₂O droplets at temperatures below 235 K (ref. 8). Nevertheless, the crystalline phase of ice that forms when aqueous droplets freeze in the atmosphere is unknown. Before this study, cubic ice (ice I_c) has been made in the laboratory using extremely fast cooling rates^{9–13}, by annealing the amorphous phase^{14,15}, and by freezing of water in nano-porous silica¹⁶. The conditions used to produce ice I_c in these previous studies are significantly different from the conditions at which ice clouds form in the Earth's atmosphere. The crystalline form of ice that precipitates in water has also been investigated theoretically. For example, thermodynamic considerations suggest that the free energy involved in forming an ice I_c nucleus are more favourable than those needed to form a nucleus of hexagonal ice (ice I_h)^{17,18}. In addition, molecular-dynamics simulations found that ice I_c crystallizes in electric fields¹⁹ and a variant of ice IX forms when the water density is 1.15 and 1.20 g cm^{–3} (ref. 20). However, the crystal structure that forms under typical atmospheric conditions is unclear from theoretical studies. In the present experimental study we investigate the phase of ice that forms when pure water and aqueous solution droplets (solute = (NH₄)₃H(SO₄)₂, HNO₃, (NH₄)₂SO₄ or NaCl) freeze homogeneously, while being cooled at rates approaching those found in the atmosphere.

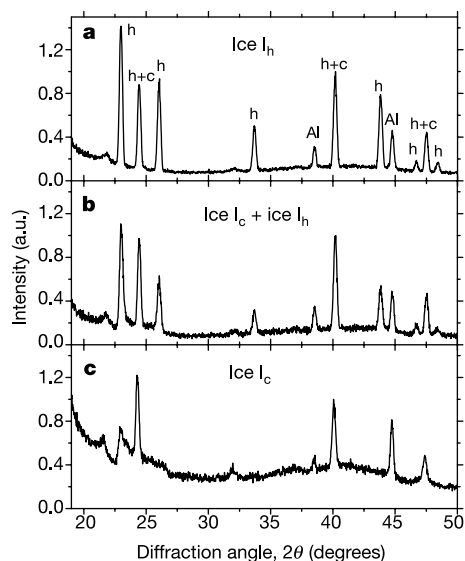


Figure 1 X-ray diffraction patterns of frozen aqueous droplets. The droplets homogeneously froze as the temperature of the cell was decreased at a rate of 10 K min^{–1}. **a**, Diffraction pattern of pure water, which froze at 235 ± 2 K and was annealed at 268 K for 35 min. **b**, Pattern for the same sample of pure water as in **a** before it was annealed. **c**, Diffraction pattern of 43 wt% (NH₄)₃H(SO₄)₂ solution droplets that froze at 188 ± 2 K. Reflections unique to ice I_h are labelled h and those common to both ice I_c and I_h are labelled h+c. Each of the diffraction patterns has been normalized to the peak at 40°. Minor peaks at 22° and 32° are reflections from the Teflon and grease. Major peaks marked 'Al' are reflections from the aluminium cooling cell.

For these experiments, water and aqueous solution droplets (between 2 and 20 μm in diameter) were suspended in an oil matrix, by emulsification, and the phase of ice that formed in the water and aqueous droplets was monitored with X-ray diffraction (see Methods). X-ray diffraction patterns of frozen pure water and aqueous solution droplets are presented in Fig. 1. Figure 1a illustrates the diffraction pattern of pure water that froze at its homogeneous freezing point ($235 \pm 2\text{ K}$) and was subsequently annealed at 268 K for 35 min. The intensities and positions of the individual reflections are in very good agreement with standard powder patterns of ice I_h (ref. 21). Figure 1b illustrates the diffraction pattern of frozen pure water droplets that were cooled directly to 173 K without annealing. The relative intensities of the reflections are clearly different. In this pattern, each of the peaks unique to ice I_h are reduced in intensity relative to those common to both ice I_c and I_h . This indicates that a significant quantity of ice I_c forms in pure water droplets.

Figure 1c illustrates the diffraction pattern of frozen solution droplets containing a 3:1 molar mixture of ammonium sulphate to sulphuric acid that froze at $188 \pm 2\text{ K}$. This mixture corresponds to the letovicite, $(\text{NH}_4)_3\text{H}(\text{SO}_4)_2$, stoichiometry. This diffraction pattern is in very good agreement with diffraction patterns of ice I_c reported in the literature^{9,10,13,14}. It is clear from this pattern that the major ice I_h reflections at 34° and 44° are completely absent. Also, the peak at 26° appears to be absent, although the baseline is uncertain in this region. The absence of these peaks shows that bulk ice I_h does not form when these solutions freeze at temperatures $<190\text{ K}$ and shows that ice I_c is the dominant product. The peak at 23° , which is usually associated with the (100) reflection of ice I_h , is present in the diffraction pattern. This feature has been observed in previous studies of ice I_c employing X-ray diffraction^{9,10,13,14} and neutron diffraction²²; it has been associated with hexagonal-like stacking faults, believed to be an intrinsic property of ice I_c (ref. 22). Inspection of the diffractograms of ice I_c reported in the references mentioned above^{9,10,13,14} reveals that the region between 22° and 27° is raised above the background, which is also the case in the present study. This broad feature is probably also related to stacking faults.

In Fig. 2 we have plotted, for the four solute types and a range of solute concentrations, the ratio of the intensity of an exclusive ice I_h peak at 44° (I_{44}) to the intensity of a peak common to both ice I_c and ice I_h at 40° (I_{40}). A ratio of 0.80 indicates that the sample froze to pure ice I_h (based on the diffraction pattern of fully annealed ice illustrated in Fig. 1a) and a ratio of zero indicates that the droplets froze exclusively to ice I_c with some stacking faults. Values between 0.80 and zero indicate the droplets froze as mixtures of ice I_c and ice I_h , with a decrease in ratio corresponding to an increase in the

amount of ice I_c . It is clear from Fig. 2 that the proportion of ice I_c exhibits a strong solute type and concentration dependence. For example, $(\text{NH}_4)_3\text{H}(\text{SO}_4)_2$ solutions always freeze to form a significant amount of ice I_c , while HNO_3 droplets can freeze to form $\sim 100\%$ ice I_h . These observations are not unprecedented: previous studies revealed complex crystallization behaviour when impurities are added to a system²³, although the mechanisms for this behaviour are not well understood. Figure 2 also clearly shows that solution droplets, of all four solutes, freeze to form close to 100% ice I_c below 190 K. Although these temperatures are low, they are in the range most important for the tropical tropopause and polar stratosphere.

In addition to examining the crystalline phase of ice that forms when aqueous droplets freeze, we also investigated the stability of ice I_c . The timescale of the cubic to hexagonal phase transition was much longer in our experiments than in many previous studies^{5,14}. The lifetime of ice I_c samples prepared by annealing vapour-deposited amorphous ice is of the order of 1 min at 230 K (refs 5, 14). In our studies, pure water droplets, which froze at 235 K, still consisted of a significant proportion of ice I_c ($I_{44}/I_{40} = 0.67 \pm 0.04$) after annealing for 10 min at 253 K. Our measurements are consistent with the long lifetimes of low-surface-area ice I_c observed by Mayer and Hallbrucker¹⁰, who created their ice films by hyperquenching liquid water droplets. Vapour-deposited ice films are known to have large surface areas²⁴, and this may increase the rate of mass transport from cubic ice to hexagonal ice crystals⁵ as well as increase the rate of nucleation of hexagonal ice on cubic ice crystals¹⁰. This would result in a shorter lifetime of cubic ice. In our experiments, mass transport (via the vapour phase) and possibly nucleation at the surface of cubic ice crystals, are blocked, as the particles are surrounded by an oil matrix. We therefore may only observe the bulk solid-to-solid phase transition.

The results presented above offer a significant insight into the crystallization of ice, a process that is poorly understood. The formation of ice I_c in pure water and aqueous solutions may suggest that water follows the Ostwald's step rule²⁵, which was first proposed in 1897, and states that the metastable crystalline phase should nucleate in preference to the stable phase. The nucleation of ice I_c is also consistent with thermodynamic considerations^{17,18}. These results should help constrain theoretical calculations and lead to a better understanding of the freezing of ice in general. Below we outline the importance of these results for Earth's atmosphere.

We clearly show that pure water droplets of approximately 10 μm diameter freeze to a significant proportion of ice I_c . One factor in determining the final proportion of ice I_c in frozen droplets is the relative rates of heat dissipation (L) and heat production (P) in the individual particles during the freezing process. If a particle heats up sufficiently, ice I_c will be annealed to ice I_h . Approximate

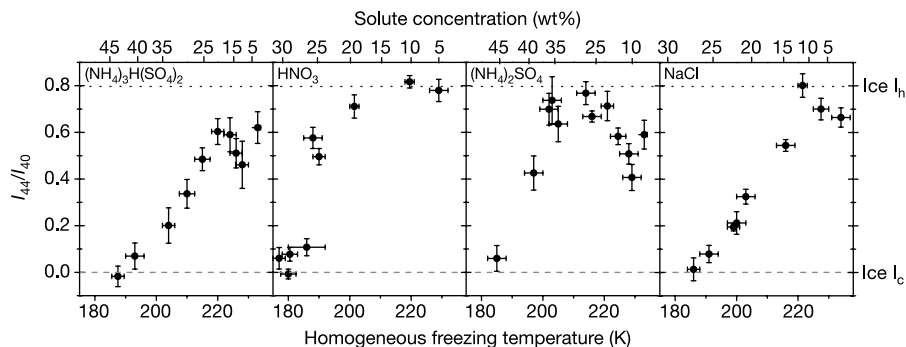


Figure 2 The intensity ratio, I_{44}/I_{40} , as a function of freezing temperature and concentration of aqueous droplets. Values of 0.80 (dotted line) correspond to pure ice I_h , while values of zero (dashed line) correspond to ice I_c in the absence of any bulk ice I_h . Results are presented for the four solutes investigated. The primary abscissa corresponds to the temperature at which each sample froze. The secondary abscissa corresponds to

the concentrations of the samples. The intensity, I , of the peaks was determined by measuring the area of the respective peaks and the uncertainty is represented by the y-error bars. The x-error bars represent the temperature range over which homogeneous freezing was observed in a single experiment, in addition to the uncertainty in the absolute temperature. Each data point represents a single freezing experiment.

calculations² suggest that when pure water particles (10 μm in diameter) freeze homogeneously in oil emulsions, the terms P and L are of a similar magnitude and the particles may not warm up sufficiently to anneal all ice I_c to ice I_h , which is consistent with our experimental results. However, the same particles in the atmosphere will lose heat to the environment roughly 10 times more slowly, and therefore we expect heating to be more significant in this case. In the atmosphere it is unlikely that a 10 μm water droplet will contain a significant amount of cubic ice after freezing, although the nucleation process may begin with a cubic ice embryo. It should also be borne in mind that L also depends on the droplet size. A smaller droplet will dissipate heat more rapidly and will therefore be more likely to contain a significant proportion of ice I_c .

In contrast to pure water and dilute solution droplets, when concentrated solution droplets freeze homogeneously in the atmosphere $L \gg P$. This is because the crystal growth rate, and therefore P , is very slow in low-temperature viscous liquids. For example, separate experiments using optical microscopy show that a 40 wt% ammonium sulphate particle (approximately 10 μm in diameter) takes approximately one third of a second to freeze at 200 K. Concentrated solution droplets will experience no appreciable heating when they freeze (at ≤ 200 K), and as a result, they will crystallize as ice I_c . Furthermore, if these droplets should heat up during freezing, their temperature may still be in a regime where the ice I_c to ice I_h transition is slow. These arguments and our experimental results strongly suggest that cubic ice will form in the atmosphere and should be considered when discussing cloud microphysics, atmospheric chemistry and the Earth's climate system. Our results also support previous speculations that cubic ice forms in the atmosphere^{3–5,9–11}.

As pointed out by Murphy⁵, the transient presence of ice I_c in the upper troposphere may modify the evolution of cirrus clouds through a process of mass transfer from ice I_c to ice I_h particles. At low temperatures ice I_c will form, and some of these particles may be converted to ice I_h by nucleation of ice I_h at their surface. (As mentioned above, such nucleation at the surface of ice I_c particles may have been blocked in our experiments by the oil matrix.) Mass transfer from ice I_c to ice I_h particles will then occur, because the vapour pressure of the former phase is larger than that of the latter ice (ice I_c is metastable to ice I_h , and therefore, ice I_c has a higher vapour pressure, although this has never been measured⁵). Under specific conditions this process may result in clouds with larger particles that have a greater sedimentation velocity, thus enhancing dehydration of the upper troposphere and tropopause region⁵.

Furthermore, the formation of ice I_c in the upper atmosphere will lead to ice clouds with a relative humidity with respect to ice I_h (RH_i) greater than 100% (ref. 5; RH_i is defined as the water vapour pressure divided by the equilibrium vapour pressure over ice I_h). Recently, Gao *et al.*²⁶ reported RH_i values of around 130% within cirrus below 203 K. Above this temperature, the RH_i values within cirrus were close to the expected 100%. The temperature range over which Gao *et al.* observed $\text{RH}_i > 100\%$ is close to the range over which we observed almost exclusive cubic ice formation. Gao *et al.* attributed the high RH_i values to a new class of HNO_3 -containing ice particles. An additional explanation is that the high RH_i values were due to the presence of cubic ice.

The formation of cubic ice in the atmosphere may also have other consequences. For example, it may have a different heterogeneous chemistry from that of hexagonal ice, as the binding of reactive species to the surface may be altered. Further detailed studies, including vapour pressure measurements, are needed to accurately assess the importance of cubic ice in the atmosphere. $\square$

Methods

Sample preparation

Water and aqueous droplets containing $(\text{NH}_4)_2\text{SO}_4$, HNO_3 , NaCl or $(\text{NH}_4)_3\text{H}(\text{SO}_4)_2$ were suspended in an oil matrix by emulsification, and X-ray diffraction was used to determine

the phase of ice that formed when the droplets froze homogeneously. The emulsified samples were prepared by mixing an aqueous solution of known composition with an oil phase (10 wt% lanolin in a hydrocarbon oil) and agitated for about 10 min (ref. 27). Lanolin was the emulsifying agent. The agitation process resulted in an emulsion of aqueous droplets between 2 and 20 μm in diameter. The emulsified samples were placed in a temperature controlled cell consisting of an aluminium base and covered with a thin film of Teflon ($\sim 30 \mu\text{m}$ thick).

Experimental procedure

X-ray diffraction measurements were performed with an X-ray diffractometer in the Bragg-Brentano reflection configuration, with $\text{Cu K}\alpha_1$ radiation, a Ge crystal monochromator and a $\text{Si}(\text{Li})$ semiconductor detector. In a typical experiment the temperature of the cell containing the emulsified sample was decreased from ambient to 173 K at a rate of 10 K min^{-1} . In order to measure the temperature at which the supercooled liquid droplets homogeneously froze to ice, the diffraction angle of a strong reflection associated with both ice I_c and I_h (diffraction angle, $2\theta = 24^\circ$) was continually monitored as the cell was cooled. Around 10 s was required to scan across this peak and during this time the temperature of the cell changed by $\sim 1.5 \text{ K}$. Freezing of droplets to ice is a stochastic process and occurs over a range of temperatures for any one sample. The x-error bars in Fig. 2 represent this range as well as the uncertainty of the temperature measurement. The measured homogeneous freezing points are typically within 3 K of previous measurements of the freezing of droplets in emulsions and on hydrophobic surfaces^{27–29} as well as a parameterization relating water activity to the homogeneous freezing temperature³⁰. Once at 173 K, the diffraction pattern of the sample between $2\theta = 19–50^\circ$ was recorded. This range covered all of the strong ice I_c and ice I_h reflections¹⁴ and selected examples of the resulting patterns are illustrated in Fig. 1.

Data analysis

From the diffraction patterns, we determined the ratio of the intensity of an exclusive ice I_h peak at 44° (I_{44}) and a peak common to both ice I_c and ice I_h at 40° (I_{40}). This ratio was then used to determine the conditions at which the droplets froze as pure ice I_h and ice I_c with stacking faults. This ratio was also used to elucidate temperature and concentration dependences (Fig. 2). When both ice I_c and ice I_h formed in a freezing experiment we did not use traditional quantitative X-ray analysis, which assumes that the peak intensities are proportional to concentrations, to quantify the relative amounts of ice I_c and ice I_h . This is because recent calculations in our group suggest that some of the peak intensities may vary in a nonlinear way with ice I_c content owing to the presence of stacking faults.

Received 7 September 2004; accepted 25 January 2005; doi:10.1038/nature03403.

- Young, K. C. *Microphysical Processes in Clouds* (Oxford Univ. Press, New York, 1993).
- Pruppacher, H. R. & Klett, J. D. *Microphysics of Clouds and Precipitation* (Kluwer, Dordrecht, 1997).
- Whalley, E. Scheiner's halo: evidence for ice I_c in the atmosphere. *Science* **211**, 389–390 (1981).
- Riikonen, M. *et al.* Halo observations provide evidence of airborne cubic ice in the Earth's atmosphere. *Appl. Opt.* **39**, 6080–6085 (2000).
- Murphy, D. M. Dehydration in cold clouds is enhanced by a transition from cubic to hexagonal ice. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL018566 (2003).
- Peter, T. Microphysics and heterogeneous chemistry of polar stratospheric clouds. *Annu. Rev. Phys. Chem.* **48**, 785–822 (1997).
- Baker, M. B. Cloud microphysics and climate. *Science* **276**, 1072–1078 (1997).
- DeMott, P. J. in *Cirrus* (eds Lynch, D. K., Sassen, K., Starr, D. C. & Stephens, G.) 102–135 (Oxford Univ. Press, New York, 2002).
- Kohl, I., Mayer, E. & Hallbrucker, A. The glassy water-cubic ice system: a comparative study by X-ray diffraction and differential scanning calorimetry. *Phys. Chem. Chem. Phys.* **2**, 1579–1586 (2000).
- Mayer, E. & Hallbrucker, A. Cubic ice from liquid water. *Nature* **325**, 601–602 (1987).
- Huang, J. F. & Bartell, L. S. Kinetics of homogeneous nucleation in the freezing of large water clusters. *J. Phys. Chem.* **99**, 3924–3931 (1995).
- Lepault, J., Bigot, D., Studer, D. & Erk, I. Freezing of aqueous specimens: an X-ray diffraction study. *J. Microsc.* **187**, 158–166 (1997).
- Dowell, L. G., Moline, S. W. & Rinfret, A. P. A low-temperature X-ray diffraction study of ice structure formed in aqueous gelatin gels. *Biochim. Biophys. Acta* **59**, 158–167 (1962).
- Dowell, L. G. & Rinfret, A. P. Low-temperature forms of ice as studied by X-ray diffraction. *Nature* **188**, 1144–1148 (1960).
- Jenniskens, P. & Blake, D. F. Structural transitions in amorphous water ice and astrophysical implications. *Science* **265**, 753–756 (1994).
- Steytler, D. C., Dore, J. C. & Wright, C. J. Neutron diffraction study of cubic ice nucleation in a porous silica network. *J. Phys. Chem.* **87**, 2458–2459 (1983).
- Takahashi, T. On the role of cubic structure in ice nucleation. *J. Cryst. Growth* **59**, 441–449 (1982).
- Kieft, H., Clouter, M. J. & Whalley, E. Cubic ice, snowflakes, and rare gas solids: surface energy, entropy, and the stability of small crystals. *J. Chem. Phys.* **81**, 1419–1420 (1984).
- Svishchev, I. M. & Kuslik, P. G. Crystallization of liquid water in a molecular dynamics simulation. *Phys. Rev. Lett.* **73**, 975–978 (1994).
- Yamada, M., Mossa, S., Stanley, H. E. & Sciortino, F. Interplay between time-temperature transformation and the liquid-liquid phase transition in water. *Phys. Rev. Lett.* **88**, 195701 (2002).
- McClune, W. F. *Powder Diffraction File-2* (The International Centre for Diffraction Data, Newtown Square, 2002).
- Kuhs, W. F., Bliss, D. V. & Finney, J. L. High-resolution neutron powder diffraction study of ice- I_c . *J. Phys. Colloq.* **48**, 631–636 (1987).
- Eidelman, N., Azoury, R. & Sarig, S. Reversal of trends in impurity effects on crystallization parameters. *J. Cryst. Growth* **74**, 1–9 (1986).
- Murray, B. J. & Plane, J. M. C. The uptake of atomic oxygen on ice films: Implications for noctilucent clouds. *Phys. Chem. Chem. Phys.* **5**, 4129–4138 (2003).
- Ostwald, W. Studien über die Bildung und Umwandlung fester Körper. *Z. Phys. Chem.* **22**, 289–330 (1897).

26. Gao, R. S. *et al.* Evidence that nitric acid increases relative humidity in low-temperature cirrus clouds. *Science* **303**, 516–520 (2004).
27. Chang, H. Y. A., Koop, T., Molina, L. T. & Molina, M. J. Phase transitions in emulsified $\text{HNO}_3/\text{H}_2\text{O}$ and $\text{HNO}_3/\text{H}_2\text{SO}_4/\text{H}_2\text{O}$ solutions. *J. Phys. Chem. A* **103**, 2673–2679 (1999).
28. Koop, T., Kapilashrami, A., Molina, L. T. & Molina, M. J. Phase transitions of sea-salt/water mixtures at low temperatures: Implications for ozone chemistry in the polar marine boundary layer. *J. Geophys. Res.* **105**, 26393–26402 (2000).
29. Bertram, A. K., Koop, T., Molina, L. T. & Molina, M. J. Ice formation in $(\text{NH}_4)_2\text{SO}_4\text{-H}_2\text{O}$ particles. *J. Phys. Chem. A* **104**, 584–588 (2000).
30. Koop, T., Luo, B. P., Tsias, A. & Peter, T. Water activity as the determinant for homogeneous ice nucleation in aqueous solutions. *Nature* **406**, 611–614 (2000).

Acknowledgements We thank D. M. Murphy for several helpful discussions on cubic ice, G. N. Patey for discussions on theoretical calculations, and A. Lam and B. Patrick for their assistance with X-ray diffraction measurements and interpretation. We are also grateful to M. Raudsepp for discussions on crystallography. This research was supported by the Canadian Foundation for Climate and Atmospheric Sciences, the Natural Sciences and Engineering Research Council of Canada, and the Canadian Foundation for Innovation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.K.B. (bertram@chem.ubc.ca).

Measuring the onset of locking in the Peru–Chile trench with GPS and acoustic measurements

Katie Gagnon¹, C. David Chadwell¹ & Edmundo Norabuena²

¹Scripps Institution of Oceanography, Marine Physical Lab, San Diego, California 92093-0205, USA

²Geodesy Department, University of Miami, Miami, Florida 33149, USA

The subduction zone off the west coast of South America marks the convergence of the oceanic Nazca plate and the continental South America plate. Nazca–South America convergence over the past 23 million years has created the 6-km-deep Peru–Chile trench, 150 km offshore. High pressure between the plates creates a locked zone, leading to deformation of the overriding plate. The surface area of this locked zone is thought to control the magnitude of co-seismic release and is limited by pressure, temperature, sediment type and fluid content¹. Here we present seafloor deformation data from the submerged South America plate obtained from a combination of Global Positioning System (GPS) receivers and acoustic transponders. We estimate that the measured horizontal surface motion perpendicular to the trench is consistent with a model having no slip along the thrust fault between 2 and 40 km depth. A tsunami in 1996, 200 km north of our site, was interpreted as being the result of an anomalously shallow interplate earthquake². Seismic coupling at shallow depths, such as we observe, may explain why co-seismic events in the Peruvian subduction zone create large tsunamis.

In 2001, two precision transponder arrays were placed on the submarine portion of the South America plate, 20 and 50 km inland from the Peru–Chile trench axis. The objective was to precisely locate seafloor reference points over a two-year period, to calculate absolute horizontal deformation and model the updip limit of the seismogenic zone.

Andean subduction alternates between flat and steep slab subduction, based on the buoyancy of the oceanic crust as a result of age and thickness^{3,4}. At 12°S, flat slab subduction causes the cold oceanic plate to extend hundreds of kilometres inland at about 100 km depth before descending. Earthquakes are created within the limits of the seismogenic zone, where converging plates are

seismically coupled along the thrust. Tichelaar and Ruff¹ seismically located the downdip limit between 48 and 53 km in the flat slab region of Chile. The downdip limit is controlled by the 350 and 450 °C isotherms, in which crustal rocks begin stable sliding and reach the brittle–ductile transition^{1,5,6}. Thermal models at various latitudes of the Chilean subduction zone show the 350 °C isotherm at a depth greater than 100 km as a result of flat subduction^{1,4}. Oleskevich showed that for colder subducting plates, 40 km was still the downdip limit because of the intersection of the thrust fault and the continental forearc Moho. Hydrous minerals in the mantle wedge, such as serpentinite and talc, change the frictional properties and enable stable sliding. The Moho intersection occurs between 40 and 45 km depth or 200 km distance from the trench at 12°S. Land GPS measurements of surface deformation have been modelled to show a downdip limit of about 50 km (refs 7, 8).

The updip limit is controlled by the thickness of sediment going into the trench and the heat flow from the oceanic lithosphere. In the upper plate, unconsolidated sediments in the accretionary prism lack the strength to support a stress until sufficient cementation and consolidation has occurred. The stable to unstable sliding transition is thermally controlled and can be caused by dewatering of smectite to illite between 100 and 150 °C, opal alteration to quartz at 100 °C, low-grade metamorphic zeolite and calcite replacement of clay between 75 and 175 °C, pressure solution and quartz cementation above 150 °C and hydrocarbon migration in accretionary prisms between 100 and 150 °C (ref. 5). Thermal models of the flat subduction portion of the Chilean trench show a 100 °C isotherm at 7–12 km depth and 50 km from the trench. Seismology has been used to locate the updip limit in Costa Rica between 8 and 20 km depth, depending on heat flow and hydrothermal circulation⁹.

A combination of GPS and acoustic travel times allow centimetre-level positioning of seafloor precision transponder arrays, a method shown in Fig. 1 (refs 10, 11). We collected 120 h of continuous 1-Hz GPS and acoustic data at the acoustic centre of two precision transponder arrays in 2001 and 2003. The amount of time required on site to obtain a repeatable solution is specific to the sea state, water properties, water depth and type of transponder used for this study. The resolution in the final position for the two arrays, for the two years, changes by less than ± 10 mm after 80 h of data collection. We also collected conductivity–temperature–depth casts to calculate the ocean sound speed profile, and tide-gauge and

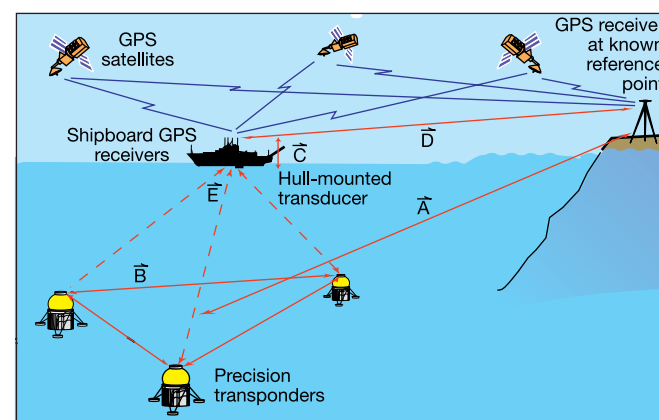


Figure 1 The GPS–acoustic approach to measure seafloor motion. Three precision transponders form an array with circle radius equal to the water depth. Relative transponder positions (**B**) are measured by acoustic ranges to a towed transducer (not shown). Dual-frequency 1-Hz GPS carrier phase data are collected at the ship and onshore (**D**). An optical survey connects phase centres from shipboard GPS antennas to the hull-mounted hydrophone (**C**). Two-way acoustic travel times are collected between the ship and transponder array (**E**). These vector components determine the horizontal components of **A**. Acoustic velocity variations do not bias the horizontal component of **A**.

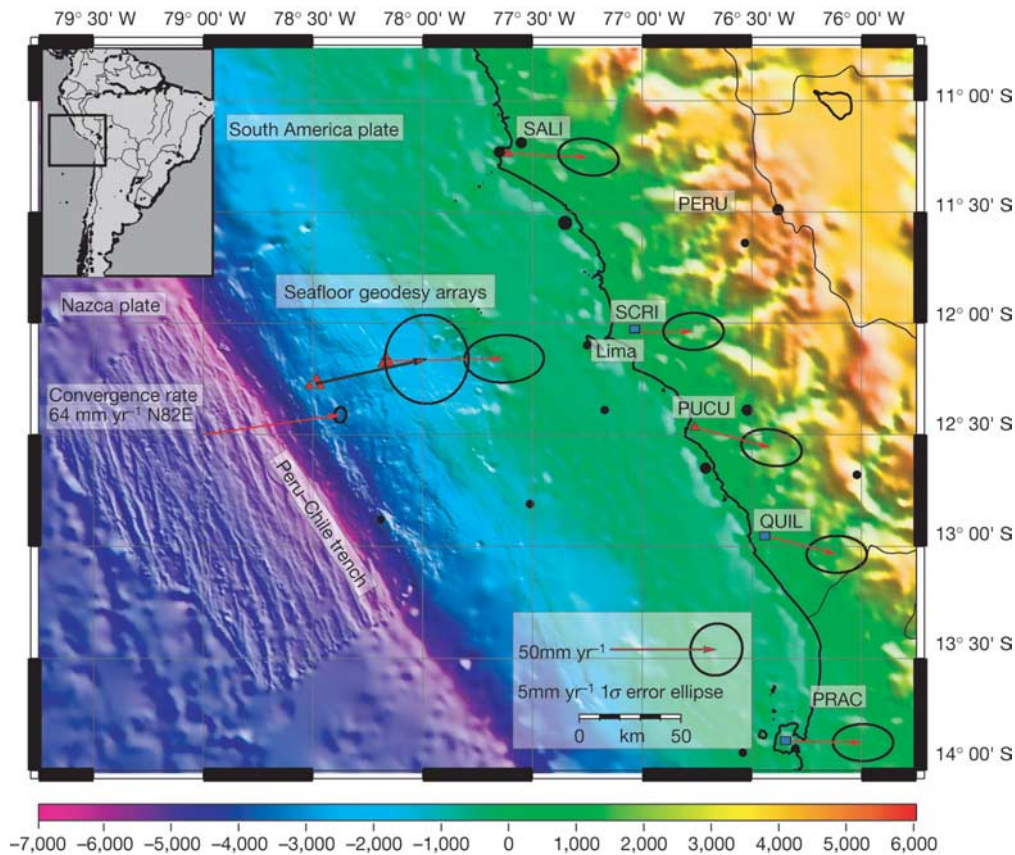


Figure 2 Bathymetric map of seafloor geodesy sites off the coast of Peru. Bathymetric data were collected in 2001 and 2003 from R/V *Roger Revelle* with SIMRAD EM 120. Circles represent earthquakes from the US Geological Survey/National Earthquake Information Center Preliminary Determinations of Epicenters solutions between data

collection periods. The REVEL plate convergence vector (64 mm yr^{-1} at $\text{N } 82^\circ \text{ E}$) is shown relative to stable South America¹⁷. Vector solutions for Nazca convergence, temporary land stations and transponder arrays are shown relative to stable South America through the use of ITRF00 (ref. 25).

pressure-sensor data to calculate the depth of each transponder below sea level. Distant land GPS stations AREQ, BOGT, GALA, SANT, KOUR and FORT defined the ITRF 2000 reference frame for South America and were used to pinpoint the positions of three temporary 1-Hz GPS stations around Lima. These, in turn, were used to define the positions of three 1-Hz shipboard GPS antennas mounted on 12-m towers. GPS data were processed with the NASA Jet Propulsion Laboratories GIPSY OASIS II software¹². Surveying techniques were employed on the ship to transfer the position of the hull-mounted transducer to the GPS antennas. The tower design and repetition of optical surveys ensure uncertainties of the order of millimetres¹³. After 100–120 h of data collection at each site in 2001 and 2003, the acoustically positioned seafloor transponders could be referenced to the motion of the stable craton.

The resulting trench perpendicular and parallel velocities of the high-rate land stations, land stations from ref. 8 and seafloor transponder arrays are shown in Fig. 2 and Table 1 (see also Supplementary Information).

An elastic dislocation model was created by using the program

3D-def from the University of Memphis, shown in Fig. 3a (ref. 14). The geometry of the thrust fault was based on refraction seismic profiles at 12° S from GEOPECO expeditions and referenced to seismic observations, Fig. 3b (refs 15, 16). The trench axis was chosen as the surface of the elastic halfspace, or 6 km below the sea surface. This will preserve the geometric and material properties of the two plates near the trench axis, the area of concern in this study. Two planes account for varying dip of the fault plane: 6° from 0–3 km and 10.5° to 60 km depth below the sea floor, or the bottom of the model profile. The displacement of elements along the fault was determined by using the components of the convergence vector perpendicular and parallel to the trench and in the downdip direction, 59 mm yr^{-1} downdip and 24 mm yr^{-1} along strike. The Sella calculation of the Nazca–South America Euler vector was used in this study for its comprehensive coverage. It was possible to compare each station velocity that defined the stable South American craton with those used in the Sella model. Euler predictions for Nazca–South America convergence range from 75 mm yr^{-1} at 81° (ref. 19), 68 mm yr^{-1} at 85° (ref. 20),

Table 1 Velocities and 1σ errors of geodetic and transponder stations

Station	Distance from trench axis (km)	Latitude	Longitude	Vector (mm yr^{-1})	Azimuth	Velocity (mm yr^{-1})	
						Trench perpendicular	Trench parallel
Convergence	–	$11^\circ 5.00' \text{ S}$	$79^\circ 0.00' \text{ W}$	64	$\text{N } 82^\circ \text{ E}$	60 ± 1	20 ± 1
SALI	150	$11^\circ 14.28' \text{ S}$	$77^\circ 36.72' \text{ W}$	37	$\text{N } 92^\circ \text{ E}$	31 ± 6	19 ± 4
SCRI	160	$12^\circ 2.40' \text{ S}$	$77^\circ 1.80' \text{ W}$	26	$\text{N } 89^\circ \text{ E}$	22 ± 5	13 ± 4
PUCU	170	$12^\circ 27.96' \text{ S}$	$76^\circ 45.54' \text{ W}$	37	$\text{N } 105^\circ \text{ E}$	26 ± 6	26 ± 4
QUIL	170	$12^\circ 57.00' \text{ S}$	$76^\circ 26.40' \text{ W}$	34	$\text{N } 105^\circ \text{ E}$	24 ± 5	24 ± 4
PRAC	130	$13^\circ 52.20' \text{ S}$	$76^\circ 21.60' \text{ W}$	37	$\text{N } 91^\circ \text{ E}$	32 ± 5	19 ± 4
Landward array	20	$12^\circ 10.09' \text{ S}$	$78^\circ 9.92' \text{ W}$	55	$\text{N } 89^\circ \text{ E}$	48 ± 7	26 ± 6
Seaward array	50	$12^\circ 16.36' \text{ S}$	$78^\circ 29.18' \text{ W}$	53	$\text{N } 78^\circ \text{ E}$	50 ± 7	15 ± 8

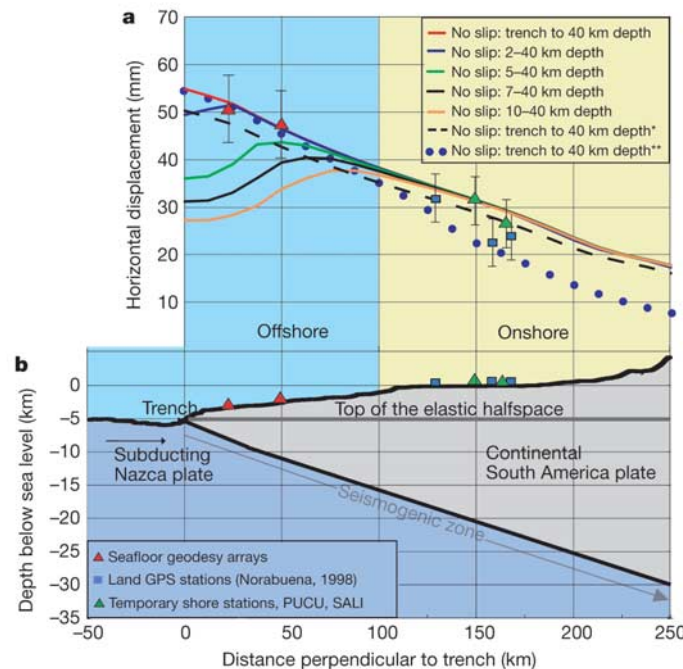


Figure 3 Models of surface deformation and plate organization. **a**, Modelling surface deformation: 3D-def model of the surface expression of a no-slip condition between the Nazca and South America plates. The onset of the no-slip zone was successively lowered from the trench to 10 km depth. Asterisk, model uses the convergence vector from ref. 20. Double asterisk, model uses a steeper dip. **b**, Plate organization: trench perpendicular

58 mm yr⁻¹ at 80° (ref. 18), 64 mm yr⁻¹ at 80° (ref. 7), to that used here, 64 mm yr⁻¹ at 82° (ref. 17). The updip limit of the locked zone was successively lowered on the fault with each model and compared with the observed trench perpendicular velocities. Two additional models differed by using the convergence rate predicted in ref. 20 (model denoted by the asterisk) and the other by using a steeper dip as has been predicted by seismological results (denoted by two asterisks).

The measured horizontal surface motion perpendicular to the trench is consistent with a model having no slip along the thrust fault between 2 and 40 km depth²¹. It is possible that the mechanics of the wedge allow the no-slip condition to extend from 2 km depth up to the trench axis, but the data are unable to resolve the fault behaviour seaward of the two arrays. The shallower geometry of the thrust fault has been well determined by recent GEOPECO seismic refraction and reflection studies. Changes to the modelled fault geometry downdip of those do not affect the interpretation of the seafloor data. In addition, varying the convergence rate does not change the interpretation of the seafloor data. The far-field vector calculation for PUCU, SALI and rover stations support previously determined downdip locking at 40 km depth^{8,22,23}.

Tsunami hazards investigated in ref. 2 present the 1996 earthquake and resulting tsunami 200 km north of our site as the result of an anomalously shallow, interplate earthquake 10 km below the sea floor. Citing examples from Nicaragua, Japan, Alaska and Peru, those authors show that the largest tsunamis are initiated by earthquakes below unconsolidated accretionary prisms. Evidence for continuous or episodic shallow locking may help to explain the "shallow extension of the seismogenic zone" that was suggested to have caused the Peruvian tsunami. The absence of slip at 2 km depth could mean that co-seismic events in the Peruvian subduction zone are more likely to create large tsunamis than subduction zones with deeper updip limits.

The updip limit can be compared with thermal models of flat slab subduction in Chile. There the convergence rate was 84 mm yr⁻¹ and the plate age 45 Myr, whereas at 12° S the convergence rate used

profile from the Nazca to the South America plate around 12° S. Topography combines multibeam sonar and satellite altimetry data²⁶. Displacement (and 1σ errors) perpendicular to the trench relative to stable South America are shown as triangles for temporary land stations and transponder arrays, and squares for land stations from ref. 8.

was 20 mm yr⁻¹ slower and the plate age 35 Myr. The locations also differ in the amount of sediment accretion. Bounded by the Atacama Desert and the Mendaña fracture zone to the north, the trench south of 11.5° S has much less sediment input. In southern Chile, the sediment is several hundred metres thick and fed from the south²⁴. Increased sediment and pore water may help to explain deeper estimates for the coupled zone in the southern portions of the trench. At 12° S, the small amount of trench sediment might increase the friction and decrease the depth of the seismogenic zone.

The deformation pattern of the defining Andean type margin has previously been confined to land-based GPS. Because the Peru–Chile trench is 150 km offshore, land-based methods are not sensitive to the activity of the plates near the trench. Seafloor geodesy, combing GPS and acoustics, and new seismic profiles revealing the upper geometry of the thrust fault, allow a previously unseen look at the initiation of a no-slip condition off the coast of Peru. The GPS/acoustic technique provides a method of measuring the convergence rate and increase in strain between plates that interact under water, which includes most plate boundaries around the world. □

Received 17 August 2004; accepted 27 January 2005; doi:10.1038/nature03412.

1. Tichelaar, B. W. & Ruff, L. J. Depth of seismic coupling along subduction zones. *J. Geophys. Res.* **98**, 2017–2037 (1993).
2. Satake, K. & Tanioka, Y. Sources of tsunami and tsunamigenic earthquakes in subduction zones. *Pure Appl. Geophys.* **154**, 467–483 (1999).
3. Cahill, T. & Isacks, I. Seismicity and shape of the subducted Nazca Plate. *J. Geophys. Res.* **97**, 17503–17529 (1992).
4. Gutscher, M. Andean subduction styles and their effect on thermal structure and interplate coupling. *J. S. Am. Earth Sci.* **15**, 3–10 (2002).
5. Moore, J. C. & Saffer, D. Updip limit of the seismogenic zone beneath the accretionary prism of southwest Japan: An effect of diagenetic to low-grade metamorphic processes and increasing effective stress. *Geology* **29**, 183–186 (2001).
6. Oleskevich, D. A., Hyndman, R. D. & Wang, K. The updip and downdip limits to great subduction earthquakes: Thermal and structural models of Cascadia, South Alaska, SW Japan, and Chile. *J. Geophys. Res.* **104**, 14965–14991 (1999).
7. Norabuena, E., Dixon, T. H., Stein, S. & Harrison, C. G. A. Decelerating Nazca–South America and Nazca–Pacific Plate motions. *Geophys. Res. Lett.* **26**, 3405–3408 (1999).
8. Norabuena, E. *et al.* Space geodetic observations of Nazca–South America convergence across the central Andes. *Science* **279**, 358–362 (1998).

9. Newman, A. V. *et al.* Along-strike variability in the seismogenic zone below Nicoya Peninsula, Costa Rica. *Geophys. Res. Lett.* **29**, 38–41 (2002).
10. Chadwell, C. D. & Bock, Y. Direct estimation of absolute precipitable water in oceanic regions by GPS tracking of a coastal buoy. *Geophys. Res. Lett.* **28**, 3701–3704 (2001).
11. Spiess, F. N. *et al.* Precise GPS/acoustic positioning of seafloor reference points for tectonic studies. *Physics Earth Planet. Inter.* **108**, 101–112 (1998).
12. Webb, F. H. & Zumberge, J. F. *An introduction to GIPSY/OASIS-II* (JPL Publication D-11088, Jet Propulsion Lab., Pasadena, California, 1997).
13. Chadwell, C. D. Shipboard towers for Global Positioning System antennas. *Ocean Eng.* **30**, 1467–1487 (2003).
14. Gombert, J. & Ellis, M. Topography and tectonics of the central New Madrid seismic zone: Results of numerical experiments using a three-dimensional boundary-element program. *J. Geophys. Res.* **99**, 20299–20310 (1994).
15. Krabbenhöft, A., Bialas, J., Kopp, H., Kukowski, N. & Hübner, C. Crustal structure of the Peruvian continental margin from wide-angle seismic studies. *Geophys. J. Int.* **159**, 749–764 (2004).
16. Hampel, A., Kukowski, N., Bialas, J., Hübner, C. & Heinbockel, R. Ridge subduction at an erosive margin: The collision of the Nazca Ridge in southern Peru. *J. Geophys. Res.* **109**, B02101, doi:10.1029/2003JB002593 (2004).
17. Sella, G., Dixon, T. & Mao, A. REVEL: A model for recent plate velocities from space geodesy. *J. Geophys. Res.* **107**, 2081, doi:10.1029/2000JB000333 (2002).
18. Angermann, D. & Klotz, J. R. Space geodetic estimation of the Nazca–South America Euler vector. *Earth Planet. Sci. Lett.* **171**, 329–334 (1999).
19. DeMets, C., Gordon, R., Argus, D. & Stein, S. Effect of recent revision to the geomagnetic reversal time scale on estimates of current plate motion. *Geophys. Res. Lett.* **21**, 2191–2194 (1994).
20. Larson, K. M., Freymueller, J. T. & Philipson, S. Global plate velocities from the Global Positioning System. *J. Geophys. Res.* **102**, 9961–9981 (1997).
21. Wang, K. & Dixon, T. “Coupling” semantics and science in earthquake research. *Eos* **85**, 180 (2004).
22. Tichelaar, B. & Ruff, L. Seismic coupling along the Chilean subduction zone. *J. Geophys. Res.* **96**, 11997–12022 (1991).
23. Bevis, M., Smalley, R. Jr, Herring, T., Godoy, J. & Galban, F. Crustal motion north and south of the Arica Deflection: Comparing recent geodetic results from the Central Andes. *Geochim. Geophys. Geosyst.* **1**, 1999GC000011 (1999).
24. Schweller, W. J., Kulm, L. D. & Prince, R. A. in *Nazca Plate: Crustal Formation and Andean Convergence* (eds Kulm, L. D., Dymond, J., Dasch, E. J., Hussong, D. M. & Roderick, R.) 323–349 (Mem. Geol. Soc. Am. 154, Geological Society of America, Boulder, Colorado, 1981).
25. Altamini, A., Sillard, P. & Boucher, C. ITRF2000: A new release of the International Terrestrial Reference Frame for earth science applications. *J. Geophys. Res.* **107**, 2214, doi:10.1029.2001JB000561 (2002).
26. Smith, W. H. F. & Sandwell, D. T. Global seafloor topography from satellite altimetry and ship depth soundings. *Science* **277**, 1957–1962 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Bevis for comments and suggestions; R. Zimmerman, D. Rimington and D. Price for engineering support; and the Captain and crew of the R/V *Roger Revelle*. We thank the Instituto Geofísico Del Perú for operating the land GPS stations and the Instituto Del Mar Del Perú, Dirección de Hidrografía y Navegación, for support at sea. This work was supported by the Marine Geology and Geophysics Program of the US National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.D.C. (cchadwell@ucsd.edu).

Cycles in fossil diversity

Robert A. Rohde & Richard A. Muller

Department of Physics and Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720, USA

It is well known that the diversity of life appears to fluctuate during the course of the Phanerozoic, the eon during which hard shells and skeletons left abundant fossils (0–542 million years ago). Here we show, using Sepkoski's compendium¹ of the first and last stratigraphic appearances of 36,380 marine genera, a strong 62 ± 3 -million-year cycle, which is particularly evident in the shorter-lived genera. The five great extinctions enumerated by Raup and Sepkoski² may be an aspect of this cycle. Because of the high statistical significance we also consider the contributions of environmental factors, and possible causes.

Sepkoski's posthumously published *Compendium of Fossil Marine Animal Genera*¹, and its earlier versions, has frequently been used in the study of biodiversity and extinction^{3,4}. For our purposes, diversity is defined as the number of distinct genera alive at any given time; that is, those whose first occurrence predates and whose last occurrence postdates that time. Because Sepkoski references only 295 stratigraphic intervals, the International Commission on Stratigraphy's 2004 time scale⁵ is used to translate the stratigraphic references into a record of diversity versus time; details are given in the Supplementary Information. Although Sepkoski's is the most extensive compilation available, it is known to be subject to certain systematic limitations due primarily to the varying availability and quality of geological sections^{6,7}. The implications of this will be discussed where appropriate.

Figure 1a shows a plot of diversity against time for all 36,380 genera in Sepkoski's *Compendium*. In Fig. 1b we show the 17,797 genera that remain when we remove those with uncertain ages (given only at epoch or period level), and those with only a single occurrence. The smooth trend curve through the data is the third-order polynomial that minimizes the variance of the difference

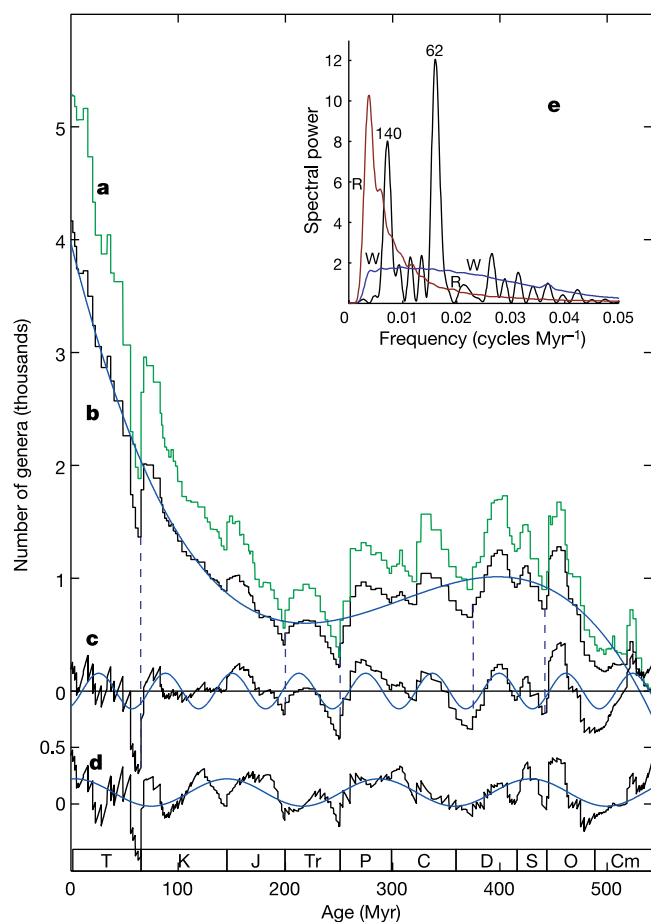


Figure 1 Genus diversity. **a**, The green plot shows the number of known marine animal genera versus time from Sepkoski's compendium¹, converted to the 2004 Geologic Time Scale⁵. **b**, The black plot shows the same data, with single occurrence and poorly dated genera removed. The trend line (blue) is a third-order polynomial fitted to the data. **c**, As **b**, with the trend subtracted and a 62-Myr sine wave superimposed. **d**, The detrended data after subtraction of the 62-Myr cycle and with a 140-Myr sine wave superimposed. Dashed vertical lines indicate the times of the five major extinctions². **e**, Fourier spectrum of **c**. Curves W (in blue) and R (in red) are estimates of spectral background. Conventional symbols for major stratigraphic periods are shown at the bottom.

between it and the data. The overall shape of Fig. 1a, b is similar to those previously published for fossil families² and for genera³. It rises rapidly at the beginning of the Phanerozoic (right side), drops to a nadir near the Permian/Triassic boundary (251 Myr ago), and then rises steeply until the present. These variations might result from evolutionary and environmental drivers⁸, observational biases⁶ or changes in the number of available geological sections⁷; for example, the sharp rise towards the present might be driven by the greater availability and study of recent sections.

Our focus will be on the short-term variations shown in Fig. 1c, obtained by subtracting the trend from Fig. 1b. The Fourier spectrum of Fig. 1c is shown in Fig. 1e. It is dominated by a strong peak with period 62 ± 3 Myr (frequency 0.0161 ± 0.0009 cycles Myr⁻¹, error is half width at half max.). The sine wave corresponding to this cycle is also shown in Fig. 1c, where it accounts for 35% of the variance. Note that because steep decreases in diversity are often followed by gradual recoveries, the peaks and valleys in the data do not precisely align with those of the sine curve. Also, some abrupt features might appear more gradual because of incomplete records⁹. We indicate the five major extinction events of Raup and Sepkoski² by dashed lines. They all occur during declining phases of the 62-Myr cycle, indicating that the 62-Myr cycle might be related to the timing or magnitude of these extinctions. However, as shown in the Supplementary Information, extinctions do not demonstrate as strong a periodicity as that shown in diversity itself.

Figure 1e has a second spectral peak, with period 140 ± 15 Myr (frequency 0.0072 ± 0.0008 cycles Myr⁻¹). To show this cycle directly in the diversity data, we subtracted the 62-Myr sine wave from Fig. 1c; the result is shown in Fig. 1d.

To determine the statistical significance of the cycles, we estimated background in two ways. In model R we assume that all diversity changes reflect a random walk, and simulate this with random permutations of the steps in Fig. 1b. Thirty thousand Monte Carlo simulations were detrended (third-order polynomial) and analysed; their average spectral power is the red line R in Fig. 1e. For model W we broke the detrended data from Fig. 1c into 20 groups and scrambled their order, thus preserving short-term correlations but randomizing the placement of major events. Thirty thousand Monte-Carlo simulations yielded the flatter blue background W. This is a more appropriate background estimate than R if the fluctuations represent perturbations about an independently driven slow trend. On the basis of these backgrounds, the probability of observing peaks at least as strong as the 62-Myr and 140-Myr spectral peaks were computed and are shown in Table 1. In doing so we considered both the significance of finding the indicated peak at the specified frequency and more generally the probability of finding a similar peak at any frequency. The 62-Myr peak is significant, with no more than a 1% chance of a similar feature occurring anywhere in the spectrum. By contrast, the 140-Myr peak can plausibly result from purely random processes, and this is likely if diversity dynamics reflect R-type behaviour. For further technical details of this computation see the Supplementary Information.

The 62-Myr cycle is also present in both first appearances

(originations) and last appearances (extinctions and pseudo-extinctions), though in both cases it is weaker. These results are included in the Supplementary Information.

The possible presence of a ~60-Myr cycle in fossil data has been suggested previously^{10–12}, but no statistical analysis was performed to confirm those qualitative observations. In fact, when we use the older time scales that were then available, we find that the 62-Myr peak would be of only questionable statistical significance. Different parts of the record shift in phase and cancel each other.

Although the 140-Myr cycle is unexpected and of ambiguous significance, we do note that it is consistent with the periods of other cycles reported in climate¹³ and cosmic rays¹⁴, and on those grounds it might merit further investigation. We also note that several authors^{15,16} have reported cycles of 26–32 Myr in diversity or extinctions. Although the 62-Myr cycle is the dominant cycle in Sepkoski's diversity data, the existence of secondary features in the middle of some cycles (Silurian, Upper Carboniferous, Lower Jurassic and Eocene) might have influenced previous reports of this ~30-Myr cyclicity.

A particularly simple way to show the appearance of the 62-Myr cycle is to separate the diversity of the 13,682 'short-lived' genera (those that endured for 45 Myr or less) from the 'long-lived' ones. This is shown, without detrending, in Fig. 2. Although the short-lived genera represent, on average, 44% of the diversity at any instant in the geological record, they are responsible for 86% of the amplitude in Fig. 1c. By contrast, the long-lived genera show few significant variations and only strongly participate in one extinction, the Permian/Triassic. We note from Figs 1c and 2a that the 62-Myr cycle is somewhat less regular and well developed during the last ~150 Myr. It is unclear whether this represents a change in the cycle or is merely an obscuring effect due to the large apparent increase in total diversity during this time.

To understand the 62-Myr cycle, it is necessary to consider environmental and causal factors that might be involved. However, given the large sampling biases affecting Sepkoski's work^{6,7}, we must also acknowledge that the cycle could be driven by a physical process affecting the fossil record, such as changes in sedimentation, rather than by a process that directly affects diversity. Another potential concern is that the durations of several geological periods (Carboniferous, Devonian and Ordovician plus Silurian) are ~60 Myr.

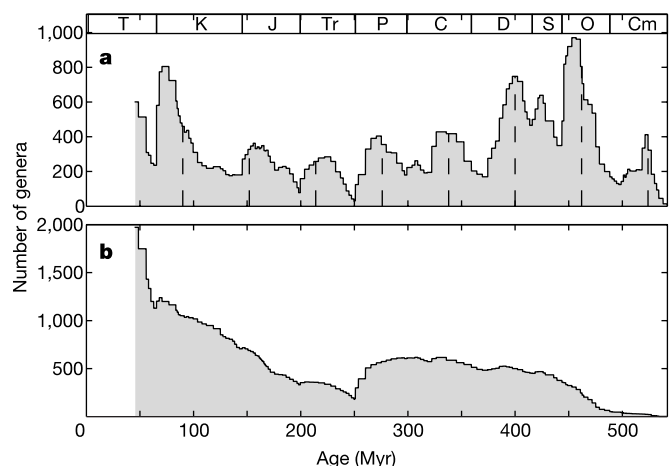


Figure 2 Diversity of short-lived and long-lived genera. This plot shows, with no detrending, the diversity of all genera that have both a first and last appearance resolved at the stage or substage level, and persisted for either 45 Myr or less (**a**) or more than 45 Myr (**b**). Genera with only single occurrences were excluded. Vertical dashed lines indicate the times of maxima of the 62-Myr sine wave of Fig. 1c. Conventional symbols for major stratigraphic periods are shown at the top.

Table 1 Likelihood of similar cycles

Probability of peaks	At this frequency		Anywhere in spectrum	
	R	W	R	W
62 Myr	$<5 \times 10^{-5}$	3.6×10^{-4}	<0.0013	0.010
140 Myr	0.12	0.0056	0.71	0.13

The probabilities of observing the 62-Myr and 140-Myr cycles, given the R or W background models as described in the paper, are shown. These are stated both in relation to finding the spectral peak in its current position and as the probability of finding a similar peak anywhere in the spectrum. For details, see the Supplementary Information.

However, most period boundaries were defined by major biological and geological events long before their ages were known. This suggests that any regularity in period durations is likely to be a response to the periodic changes influencing diversity rather than its cause.

To help understand the cycles, we examined the following seven geophysical records. First, a strong ~ 135 -Myr cycle in $\delta^{18}\text{O}$, a proxy for glaciation and climate, has been found¹³, and is statistically indistinguishable from our 140 ± 15 -Myr cycle. Cold periods in $\delta^{18}\text{O}$ precede our 140-Myr diversity maxima by 20–25 Myr. A 140-Myr cycle has been reported in other glacial indicators, although the results are disputed^{14,17–18}. Second, major volcanism is frequently invoked as a possible cause of mass extinctions. Wavelet analysis on the ages of large igneous provinces shows a minor feature near 60 Myr and no cycle near 140 Myr (ref. 19). Third, $\delta^{13}\text{C}$ is a proxy for biomass; however, spectral analysis shows no significant 62-Myr or 140-Myr cycle²⁰. Fourth, sea level changes²¹ have been associated with extinctions²². Extreme sea level states correspond to some extremes in our 62-Myr cycle, but not in a consistent way. Spectra of sea level records²¹ show peaks near the 62-Myr and 140-Myr periods, but with low statistical significance. (The Exxon sea level data are available at http://hydro.geosc.psu.edu/Sed_html/exxon.sea.) Fifth, ages of impact craters on Earth²³ show no significant periodicity at 62 Myr or 140 Myr, although it is widely believed that many craters of moderate size have never been discovered. Sixth, records of the number of explored geological formations⁷ show no 62-Myr or 140-Myr periodicity, although poor resolution could obscure such cycles. Last, a 143 ± 10 -Myr cycle in meteorite ages has been reported and attributed to variation in cosmic ray flux¹⁴.

Our searches have found no compelling match for the 62-Myr cycle; however, incomplete records and errors in time scales can obscure true periodicity. The match of our 140-Myr diversity cycle to the climate/glacial/cosmic-ray cycle might be real, because having a peak at exactly that frequency is still relatively unlikely; however, less ambiguous data would be needed to establish the connection and explain the relatively large lag between the diversity and climate cycles.

The 62-Myr cycle is strong. It might be a largely biological process or a variation in the integrity of the fossil record; however, in either case it is also worth considering geophysical processes that could be driving it. We consider seven possibilities. First, periodic passage of the solar system through molecular clouds, Galactic arms or some other structure could periodically perturb the Oort cloud and cause variations in the rate of comet impacts on the Earth²⁴. It has been argued¹⁴ that a 140-Myr period between spiral arm crossings is consistent with existing astrophysical constraints, and that such passages can affect climate by varying the cosmic ray flux. Second, laboratory simulations of mantle plumes, under idealized conditions, show relaxation oscillator modes in which plumes reach the surface at regular intervals for six to nine cycles²⁵. Similar behaviour in the Earth could cause periodic volcanism. Third, the Sun currently oscillates up and down across the Galactic plane every 52–74 Myr (ref. 26), but plausible responses²⁴ would seem to occur every mid-plane crossing (namely 26–37 Myr). Moreover, the period is not constant, but decreases to half when we encounter higher-density Galactic arms. Fourth, solar cycles could affect climate, but solar theory²⁷ predicts that long-period oscillations do not occur. Fifth, Earth orbital oscillations could affect climate. Using an orbital integration package²⁸ and nine point-mass planets, we found no significant cycles with periods of 62 Myr or 140 Myr. Changes in obliquity were not included in our calculations. Sixth, one or more companion stars to the Sun could trigger periodic comet showers. However, a 62-Myr orbit is unstable to perturbations from passing stars. The interaction of two or more short-period companions could generate a longer periodicity (for example, through beats), but our simulations suggest that mutual

perturbations would probably destroy any regularity. Last, ‘Planet X’ is a proposed large planet that perturbs the Kuiper belt and could yield periodic comet showers on the right time scales²⁹. No evidence for it exists.

Although no explanation exists, the 62-Myr cycle is not a subtle signal. It is evident even in the raw data (Fig. 1a), dominant in the short-lived genera (Fig. 2) and strongly confirmed by statistical analysis. We do not know whether this cycle is a variation in true diversity or only in observed diversity, but either case requires explanation and implies that an unknown periodic process has been having a significant impact on Earth’s environment throughout the Phanerozoic. Most models seem to make testable predictions, so we are hopeful that the cause of this behaviour will not remain a mystery for long. □

Received 11 October; accepted 23 December 2004; doi:10.1038/nature03339.

1. Sepkoski, J. A *Compendium of Fossil Marine Animal Genera* (eds Jablonski, D. & Foote, M.) *Bull. Am. Paleontol.* no. 363 (Paleontological Research Institution, Ithaca, 2002).
2. Raup, D. & Sepkoski, J. Mass extinctions in the marine fossil record. *Science* **215**, 1501–1503 (1982).
3. Sepkoski, J. in *Global Events and Event Stratigraphy* (ed. Walliser, O.) 35–61 (Springer-Verlag, Berlin, 1996).
4. Miller, A. Biotic transitions in global marine diversity. *Science* **281**, 1157–1160 (1998).
5. Gradstein, F., Ogg, J. & Smith, A. A *Geologic Time Scale 2004* (Cambridge Univ. Press, Cambridge, 2005).
6. Alroy, J. et al. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proc. Natl Acad. Sci. USA* **98**, 6261–6266 (2001).
7. Peters, E. S. & Foote, M. Biodiversity in the Phanerozoic: a reinterpretation. *Paleobiology* **27**, 583–601 (2001).
8. Jablonski, D., Roy, K., Valentine, J., Price, R. & Anderson, P. The impact of the pull of the recent on the history of marine diversity. *Science* **300**, 1133–1135 (2003).
9. Signor, P. & Lipps, J. in *Geologic Implications of Impacts of Large Asteroids and Comets on the Earth* (eds Silver, I. & Silver, P.) 291–296 (Geol. Soc. Am. Special Paper 190, Boulder, Colorado, 1982).
10. Thomson, K. S. Explanation of large scale extinctions of lower vertebrates. *Nature* **261**, 578–580 (1976).
11. Thomson, K. S. in *Patterns of Evolution as Illustrated by the Fossil Record* (ed. Hallam, A.) (Developments in Palaeontology and Stratigraphy Vol 5) 377–404 (Elsevier Scientific Publishing Company, Amsterdam, 1977).
12. Ager, D. V. The nature of the fossil record. *Proc. Geol. Assoc.* **87**, 131–159 (1977).
13. Veizer, J., Godderis, Y. & Francois, L. Evidence for decoupling of atmospheric CO₂ and global climate during the Phanerozoic eon. *Nature* **408**, 698–701 (2000).
14. Shaviv, N. Cosmic ray diffusion from the galactic spiral arms, iron meteorites, and a possible, climatic connection. *Phys. Rev. Lett.* **89**, 51102-1 (2002).
15. Fischer, A. G. & Arthur, M. A. in *Deepwater Carbonate Environments* Spec. Publ. 25 (eds Cook, H. E. & Enos, P.) 19–50 (Society of Economic Paleontologists and Mineralogists, Tulsa, Oklahoma, 1977).
16. Sepkoski, J. J. in *Global Catastrophes in Earth History* Spec. Paper 247 (eds Sharpton, V. L. & Ward, P. D.) 33–44 (Geological Society of America, Boulder, Colorado, 1990).
17. Shaviv, N. & Veizer, J. Celestial driver of Phanerozoic climate. *GSA Today* **13**, 4–10 (2003).
18. Rahmstorf, S. et al. Cosmic rays, carbon dioxide and climate. *Eos* **85**, 38–41 (2004).
19. Prokoph, A., Ernst, R. & Buchan, K. Time-series analysis of large igneous provinces: 3500 Ma to present. *J. Geol.* **112**, 1–22 (2004).
20. Veizer, J. et al. $^{87}\text{Sr}/^{86}\text{Sr}$, $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ evolution of Phanerozoic seawater. *Chem. Geol.* **161**, 59–88 (1999).
21. Haq, B., Hardenbol, J. & Vail, P. Chronology of fluctuating sea levels since the Triassic. *Science* **235**, 1156–1167 (1987).
22. Hallam, A. & Wignall, P. Mass extinctions and sea-level changes. *Earth Sci. Rev.* **48**, 217–250 (1999).
23. Whitehead, J. (<http://www.unb.ca/passc/ImpactDatabase/>) (2004).
24. Matese, J. J., Innanen, K. A. & Valtonen, M. J. in *Collisional Processes in the Solar System Astrophysics*. Space Lib. Vol. 261 (eds Marov, M. & Rickman, H.) 91–102 (Kluwer Academic Publishers, Norwell, Massachusetts, 2001).
25. Schaeffer, N. & Manga, M. Interaction of rising and sinking mantle plumes. *Geophys. Res. Lett.* **28**, 455–458 (2001).
26. Bahcall, J. & Bahcall, S. The Sun’s motion perpendicular to the galactic plane. *Nature* **316**, 706–708 (1985).
27. Bahcall, J. N. *Neutrino Astrophysics* (Cambridge Univ. Press, Cambridge, 1989).
28. Levison, H. F. & Duncan, M. J. The long-term dynamical behavior of short-period comets. *Icarus* **108**, 18–36 (1994).
29. Whitmire, D. & Matese, J. Periodic comet showers and Planet X. *Nature* **313**, 36–38 (1985).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Levine, W. Alvarez, D. Shimabukuro, R. Ernst, M. Manga and D. O’Connor for discussions and suggestions. This work was supported by the Ann and Gordon Getty Foundation and the Folger Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A.M. (ramuller@lbl.gov).

Agricultural runoff fuels large phytoplankton blooms in vulnerable areas of the ocean

J. Michael Beman¹, Kevin R. Arrigo² & Pamela A. Matson^{1,3}

¹Department of Geological and Environmental Sciences, ²Department of Geophysics and ³School of Earth Sciences, Stanford University, Stanford, California 94305, USA

Biological productivity in most of the world's oceans is controlled by the supply of nutrients to surface waters. The relative balance between supply and removal of nutrients—including nitrogen, iron and phosphorus—determines which nutrient limits phytoplankton growth. Although nitrogen limits productivity in much of the ocean^{1,2}, large portions of the tropics and subtropics are defined by extreme nitrogen depletion. In these regions, microbial denitrification removes biologically available forms of nitrogen from the water column, producing substantial deficits relative to other nutrients^{3–5}. Here we demonstrate that nitrogen-deficient areas of the tropical and subtropical oceans are acutely vulnerable to nitrogen pollution. Despite naturally high nutrient concentrations and productivity^{6–8}, nitrogen-rich agricultural runoff fuels large (54–577 km²) phytoplankton blooms in the Gulf of California. Runoff exerts a strong and consistent influence on biological processes, in 80% of cases stimulating blooms within days of fertilization and irrigation of agricultural fields. We project that by the year 2050, 27–59% of all nitrogen fertilizer will be applied in developing regions located upstream of nitrogen-deficient marine ecosystems. Our findings highlight the present and future vulnerability of these ecosystems to agricultural runoff.

Throughout large areas of the ocean, nitrogen (N) deficits characterize regions of active marine microbial denitrification^{3–5}. In these areas, biological productivity generates large amounts of organic material, the decomposition of which depletes water column oxygen concentrations, resulting in a low-oxygen environment that is favourable to denitrification. The Gulf of California (GOC, Fig. 1) contains some of the highest nutrient concentrations in the oceans⁷ and sustains highly elevated rates of biological productivity⁶. Supplied by wind-driven upwelling⁸ and inflow of nutrient-rich waters at depth⁹, concentrations of nitrogen and phosphorus (P) are well-correlated, and found in proportions corresponding to the Redfield N:P ratio of 16:1 (ref. 7). As phytoplankton growth reduces available N concentrations to near zero, an excess of nearly 1 μM P remains, suggesting a deficit of 16 μM N relative to P (ref. 7).

Although N deficits may increase vulnerability to exogenous sources of N, how the GOC and other N-deficient areas of the ocean respond to anthropogenic nitrogen pollution is largely unknown. In marine ecosystems, N pollution is associated with increased phytoplankton biomass and rates of primary production, the formation of harmful (often toxic) algal blooms, and low-oxygen conditions¹⁰. Inarguably, the effects of marine N pollution are becoming increasingly widespread and severe as a consequence of the global expansion of industrialized agriculture and the intensification of current practices. N-based agricultural fertilizers are the primary source of N pollution, and their use is predicted to double or triple over the next 50 years¹¹.

Fertilizer application rates are extremely high (250 kg N ha⁻¹, ref. 12) in the Yaqui Valley, a 2.25 $\times 10^5$ ha agricultural system located along the east coast of the GOC (Fig. 1). Surface water runoff from the Yaqui Valley is determined almost entirely by irrigation of agricultural fields, and results in the loss of large pulses of N to the

atmosphere, ground water and surface waters of the valley during irrigation events^{13–16}. Nearly 66% of the N applied in fertilizer is lost from the Yaqui Valley, 20–40% of which is lost in surface waters^{14,16}. Irrigation events are synchronized throughout the valley, typically occurring in late November, late January, early March and early April of each year. Allocation of water over a period of days results in a distinct irrigation peak for each event (Fig. 2a). Following each irrigation event, runoff flows through a network of natural and artificial drainage canals to coastal waters. Although the rate of drainage from the Yaqui Valley can be influenced by both watershed characteristics and management practice, runoff flows into the GOC within a few days of peak irrigation. In a few cases, when sufficient irrigation occurs in the days before the irrigation peak, runoff into the GOC precedes the peak; for the most part, however, irrigation waters arrive in the GOC on the day of peak irrigation or slightly after. For example, data from a conductivity–temperature–depth (CTD) recorder moored in the mouth of Bahía del Tóbari, an estuary located along the south coast of the Yaqui Valley, show a large drop in salinity and a smaller drop in temperature coincident with the peak irrigation date for January 2000 (E. Cruz-Colin, unpublished data).

To determine changes in phytoplankton biomass in the GOC driven by introduction of N in these pulses of runoff, chlorophyll *a* (Chl) data from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) were evaluated over a 5-year period. Arid conditions over the GOC allow generation of a remarkably complete time series of SeaWiFS data, with relatively few data gaps caused by cloud cover. Data extracted from a transect extending 100 km from Bahía del Tóbari across the GOC (Fig. 1), showed a strong annual cycle in phytoplankton biomass, typical of that seen throughout the GOC¹⁷. Chl concentrations were consistently highest in winter and spring, and lower in the summer months (Fig. 2b). These seasonal shifts were mirrored by changes in sea surface temperature (SST), measured using the Advanced Very High Resolution Radiometer (AVHRR); cooler SSTs in winter are associated with higher Chl levels, and vice versa (Fig. 2c).

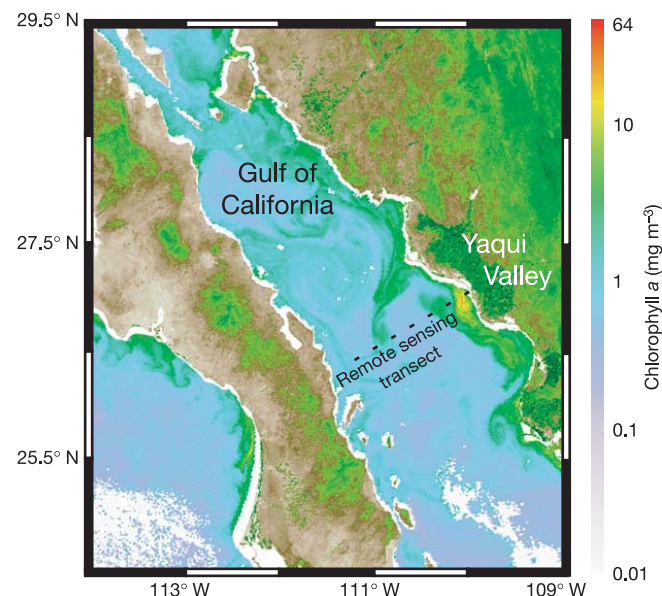


Figure 1 SeaWiFS image of chlorophyll *a* in the GOC from 6 April 1998, one day after peak irrigation. Location of remote sensing transect is shown crossing an intense phytoplankton bloom. As indicated by the colour scale, Chl concentrations in the bloom are significantly higher than elsewhere in the GOC. On land, the productive agricultural fields of the Yaqui Valley are clearly visible in a MODIS-Aqua vegetation image (NDVI) from 4 April 2003, overlaid on ocean colour data.

Superimposed on this annual cycle were rapid increases in algal biomass—discrete phytoplankton blooms—clearly identifiable throughout the 5-year Chl record. These blooms were often intense, with Chl concentrations of $>10 \text{ mg m}^{-3}$ (Figs 1 and 2b), and also quite extensive, reaching nearly 100 km across the GOC (Supplementary Video 1). Upwelling was a strong driver of many of these blooms, represented in the SST record by rapid drops in SST, as colder, nutrient-rich, deep waters were brought to the surface (Fig. 2c). Many other phytoplankton blooms, however, were associated with irrigation events in the Yaqui Valley (Fig. 2a, b). In fact, using Spearman's ranked correlation we find that periods of greater irrigation are significantly correlated with high Chl levels ($r_s = 0.47$, $P < 0.001$), despite hydrological and biogeochemical variability in the Yaqui Valley system that might mask this correlation. On the basis of this relationship, we suggest that up to 22% of the annual Chl variability in the GOC is related to N runoff from the Yaqui Valley.

Because remotely sensed SSTs are well correlated with nutrient concentrations over time, and inferred nutrient distributions are related to Chl concentrations¹⁸, a generalized linear model (GLM¹⁹) based on the relationship between SST and Chl was used to model natural Chl variability over time. Although the predictive ability of the model is limited by the strength of the SST–Chl relationship ($r^2 = 0.45$, $P < 0.01$), the model successfully removes seasonal Chl variability and winter and summer upwelling events (Fig. 2d), such as those observed in January 1998, January 1999 and August 2000. Residual Chl peaks (where Chl concentrations are greater than those predicted by the GLM) are persistent in autumn 1999 and in late spring and early summer 2001 and 2002. A similar decoupling between SST and Chl has been noted during summer in the GOC²⁰, and was attributed to rapid heating of surface waters after upwelling, leaving a remnant Chl signal of upwelling but no SST signal.

In addition, there is a strong association between residual Chl

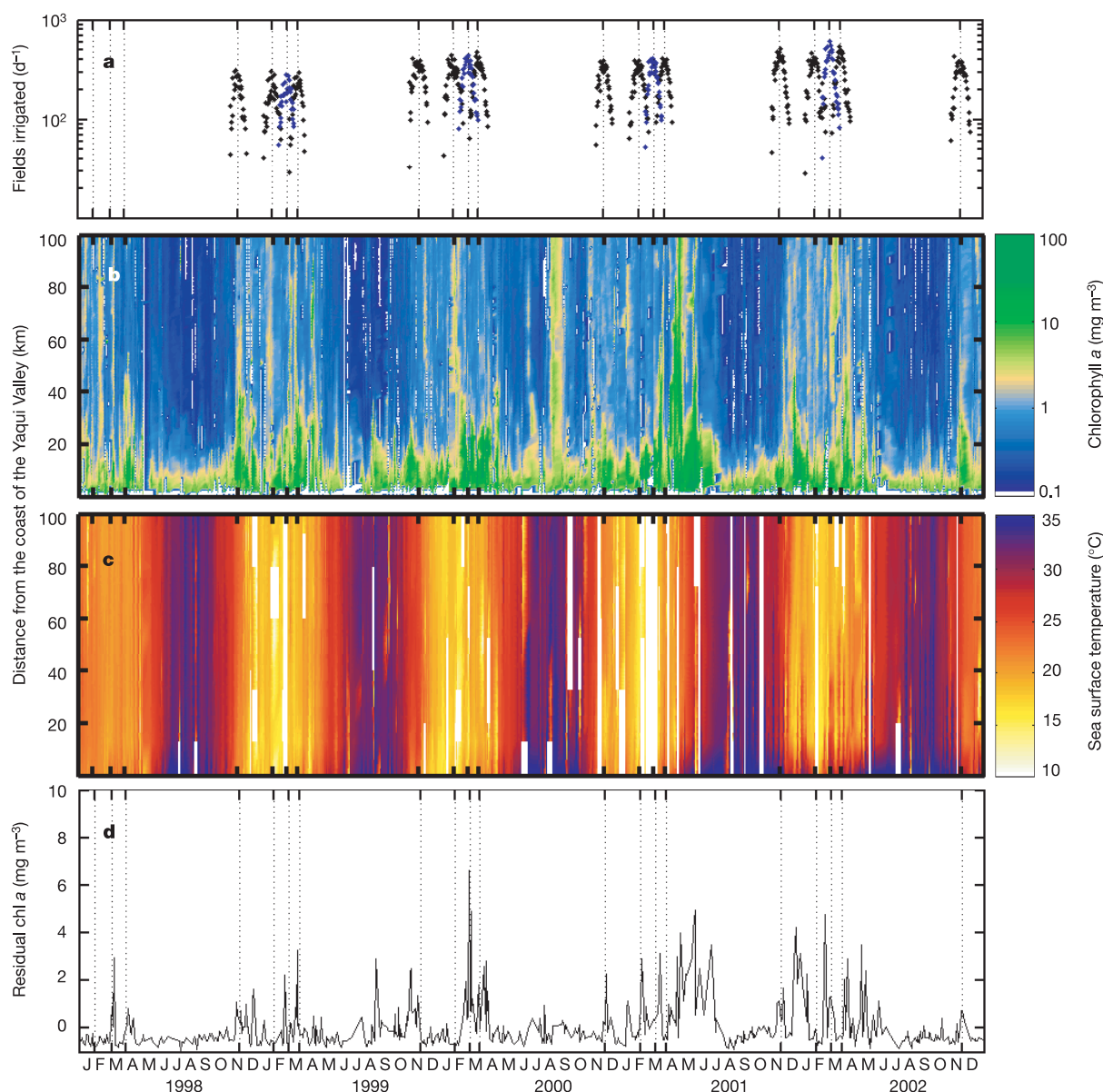


Figure 2 Five-year time series. **a**, Irrigation allotment data. **b**, SeaWiFS chlorophyll *a* data. **c**, AVHRR SST data. **d**, Residual Chl values from best-fit GLM. Horizontal axis represents time from 1998–2002; calculated peak irrigation periods are denoted by

dotted lines and tick marks along the horizontal axis. Vertical axis in **b** and **c** represents distance across transect, colour bar on right shows corresponding Chl concentrations and SSTs.

Table 1 N-deficient marine regions and fertilizer use in adjacent developing agricultural regions

N-deficient marine region	Maximum N deficit (ref. no.)	Adjacent developing agricultural region	N fertilizer use (10^6 megatonnes)		
			Value in 2000	Forecast 2020	Forecast 2050
Gulf of California	$16 \mu\text{mol l}^{-1}$ (7)	Tropical Americas	2.44	3.66	5.34
Eastern tropical Pacific	$12 \mu\text{mol kg}^{-1}$ (4,5)				
Benguela upwelling system	$40 \mu\text{mol kg}^{-1}$ (23)	Western Africa	0.246	0.469	0.711
Arabian Sea	$12 \mu\text{mol kg}^{-1}$ (24)	South Asia	14.4	21.9	33.9
Bay of Bengal	$81 \mu\text{mol l}^{-1}$ (25)				
South China Sea	$1.2 \mu\text{mol l}^{-1}$ (26)	Southeast Asia (Excluding China)	27.4	45.6	69.9
			5.22	7.52	11.7
World total			80.9	129	187

peaks in winter and early spring and irrigation events in the Yaqui Valley (Fig. 2d). Of the 20 irrigation events in the Yaqui Valley from 1998–2002, peaks in both the unfiltered Chl record and the residual Chl record are directly associated with ten of them (Supplementary Fig. S1). Peaks lag the November 2000, April and November 2001, and April 2002 irrigation events by 3–5 days, bracket the March 1999 event (occurring 3–5 days before and after) and precede the November 1998 event by about 3 days. This correspondence between Yaqui Valley irrigation events and both unfiltered and residual Chl peaks strongly suggests that phytoplankton blooms in the GOC are stimulated by agricultural runoff from the Yaqui Valley. This association is particularly strong at the beginning and end of the crop cycle in November and April, when blooms consistently follow peak irrigation in the Yaqui Valley. It is less consistent in January (blooms in 2001 and 2002 were associated with runoff but those in 1998–2000 were not), perhaps because less N was lost from the valley during these irrigation events.

In general, the amount of N lost from the Yaqui Valley during irrigation events is sufficient to balance the N requirements of associated phytoplankton blooms. Associated blooms range in size from the 54 km^2 bloom in November 2002 to the 577 km^2 bloom in March 2000 (determined using SeaWiFS data), and estimates of annual N loss from the Yaqui Valley range from 147×10^6 moles of N (based on a limited number of measurements in drainage canals¹³) to 804×10^6 moles of N (based on N application rates of 250 kg N ha^{-1} to $2.25 \times 10^5 \text{ ha}$ of fields, with 20% N loss in surface waters¹⁶). If these losses are distributed equally over the four irrigation events every year, 36.8×10^6 to 201×10^6 moles of N are lost per irrigation event. Dividing these losses by a N deficit of $16 \mu\text{M}$, and assuming a mixed layer depth of 16–46 m (ref. 21), Yaqui Valley N losses during irrigation can support phytoplankton blooms in the GOC of $50\text{--}785 \text{ km}^2$; this is in close agreement with bloom areas determined using SeaWiFS data. Although the supply of other nutrients such as P or iron (Fe) in runoff could potentially drive phytoplankton blooms in the GOC, particularly by increasing N-fixation rates²², we consider it unlikely. Phosphorus is found in excess throughout the GOC water column, with surface water concentrations $>1 \mu\text{M}$ (ref. 7), and Fe concentrations in Yaqui Valley drainage canals are low¹³. Because N losses occur in large pulses and in large part as ammonium¹³, we suspect that they effectively overwhelm the N-removal capacity of coastal systems², flow into the strongly N-limited GOC and fuel the phytoplankton blooms observed in ocean colour imagery (Fig. 1 and Supplementary Video 1).

Large portions of the tropics and subtropics may be similarly vulnerable to N in agricultural runoff, as significant N depletion (via denitrification) occurs throughout the tropical and subtropical oceans, including the eastern tropical Pacific^{3–5}, the Benguela upwelling system²³, the Arabian Sea²⁴ and the Bay of Bengal²⁵ (Table 1). Nitrogen deficits are lower in the South China Sea, but

nutrient concentrations are low and N is strongly limiting during parts of the year²⁶. These marine regions are located adjacent to rapidly developing agricultural areas in the tropical Americas, West Africa, South Asia, and Southeast Asia that will experience large increases in fertilizer use over the next few decades (Table 1). These increases outpace global trends (ref. 11, Table 1) and reflect a shift towards greater fertilizer consumption at low latitudes. In 2050, at least 27% of N fertilizer will be applied in regions adjacent to N-vulnerable marine ecosystems. Including China, this number more than doubles to 59% (Table 1). This represents a quantity greater than all N fertilizer currently applied worldwide.

Although only some portion of this applied N will reach the ocean, intensive agricultural systems like that in the Yaqui Valley are expected to develop in these regions over time²⁷, and are likely to lead to similarly high N losses. In areas where runoff from these systems reaches N-vulnerable regions, as it does in the GOC, it represents a significant new source of N and strongly influences marine ecosystem processes. The downstream effects of runoff-driven phytoplankton blooms in the GOC are not yet known, but combined with fishing pressures²⁸, carbon and energy become concentrated at lower trophic levels, with potentially significant effects for ecosystem structure and function. □

Methods

Irrigation data and timing

To determine peak irrigation dates, we fitted second-degree polynomials using least squares to irrigation allotment data from the Yaqui Valley Irrigation district. For these fits, r^2 values were between 0.74 (January 2000) and 0.91 (April 2001). We calculated the peak irrigation date as the zero value of the first derivative of the polynomial. As data from early 1998 were not available, we instead used the best available estimate of peak irrigation from the Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT, the International Wheat and Corn Improvement Center): days 31, 66 and 95 of 1998 (I. Ortiz-Monasterio, personal communication).

Remote sensing data

Information regarding procurement and processing of SeaWiFS ocean colour data and AVHRR SST data can be found in the Supplementary Information.

Generalized linear models

Time series data for use in the generalized linear model¹⁹ were taken from imagery of the GOC off the coast of the Yaqui Valley. In order to capture the full spatial variability in bloom activity, we averaged values over a 30-km transect located 10 km off the coast of the Yaqui Valley. This transect is nearly perpendicular to the previous transect, allowing for coverage of all areas of input to the GOC from the southern half of the Yaqui Valley. We report the results of our 'best-fit' model, defined as that model which was most successful in reproducing Chl variability based on variability in SST, thereby minimizing residual Chl values. The best-fit model used a log-link function, and assumes the variables are gamma-distributed. The model takes the form $\ln(\text{Chl}) = \beta_0 + \beta_1(\text{SST})$, where $\beta_0 = 3.41$ and $\beta_1 = -0.0608$. Model fitting was performed in glmab (MATLAB). Pearson residual values were calculated by subtracting modelled values from remotely sensed Chl data and scaling by the estimated standard deviation.

Irrigation–bloom lead–lag analysis

To calculate the lead and lag times between irrigation peaks and Chl peaks, we fitted a series of smoothing splines to both unfiltered Chl data and residual Chl data. Smoothing

parameters of 0.25, 0.5 and 0.75 were used for each record, generating a total of 6 curve fits. We evaluated the first derivative of these fits at an interval of 3.125 days, the mean interval between satellite images. For each of the 6 derivatives, we defined two levels of sensitivity and identified peaks that exceeded these thresholds. For the 3 unfiltered Chl curves, sensitivities of 1 and 2 mg Chl m⁻³ per 3.125-day interval were used; for the GLM residual Chl curves, sensitivities of 0.1 and 0.2 were used for all but the least smooth (parameter = 0.75) record, where sensitivities were 0.15 and 0.25. The zero value following each threshold-exceeding peak in the derivative was defined as a bloom event. Each of these 12 models identified between 20 and 92 bloom events, with significant overlap resulting in identification of 121 total bloom events (Supplementary Fig. S1). Identified blooms that occurred within a sampling interval of 3.125 days centred on irrigation peaks were considered to occur concurrently with irrigation; lead and lag timing was calculated relative to this. All statistical analyses were performed in MATLAB.

Nitrogen deficit calculations

Nitrogen deficits were reported in the literature for the Eastern Tropical Pacific^{4,5}, the Benguela upwelling system²³ and the Arabian Sea²⁴. For the GOC, Bay of Bengal and South China Sea, deficits were calculated using the ΔN formulation discussed previously²³ and based on data reported in the literature (GOC⁷, Bay of Bengal²⁵ and South China Sea²⁶).

Fertilizer data and projections

Data on use of N-based fertilizers was obtained from the United Nations Food and Agriculture Organization (FAO) Statistical Databases (<http://apps.fao.org>). World data includes all countries reporting to the FAO, and developing agricultural regions were defined as follows. Tropical Americas: Chile, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Mexico, Nicaragua, Panama and Peru; Western Africa: Angola, Benin, Burkina Faso, Cameroon, Cape Verde, Côte d'Ivoire, Democratic Republic of Congo, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Mali, Mauritania, Namibia, Niger, Nigeria, Republic of Congo, Senegal, Sierra Leone and Togo; South Asia: Bangladesh, Bhutan, India, Nepal, Pakistan and Sri Lanka; Southeast Asia: Brunei Darussalam, Cambodia, China, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam. Using a previously described approach¹¹, we found strong linear increases in fertilizer use over time in all regions, with r^2 values between 0.90 (Western Africa) and 0.97 (Southeast Asia). We projected these relationships forward to 2020 and 2050 to calculate fertilizer use.

Received 1 November 2004; accepted 17 January 2005; doi:10.1038/nature03370.

1. Tyrrell, T. The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature* **400**, 525–531 (1999).
2. Corredor, J. E., Howarth, R. W., Twilley, R. R. & Morell, J. M. Nitrogen cycling and anthropogenic impact in the tropical interamerican seas. *Biogeochemistry* **46**, 163–178 (1999).
3. Gruber, N. & Sarmiento, J. L. Global patterns of marine nitrogen fixation and denitrification. *Glob. Biogeochem. Cycles* **11**, 235–266 (1997).
4. Deutsch, C., Gruber, N., Key, R. M., Sarmiento, J. L. & Ganachaud, A. Denitrification and N₂ fixation in the Pacific Ocean. *Glob. Biogeochem. Cycles* **15**, 483–506 (2001).
5. Tyrrell, T. & Law, C. S. Low nitrate-phosphate ocean ratios corrected. *Nature* **393**, 318 (1998).
6. Zeitzschel, B. Primary productivity in the Gulf of California. *Mar. Biol.* **3**, 201–207 (1969).
7. Alvarez-Borrego, S., Rivera, J. A., Gaxiola-Castro, G., Acosta-Ruiz, M. & Schwartzlose, R. A. Nutrientes en el Golfo de California. *Ciencias Marinas* **5**, 53–71 (1978).
8. Alvarez-Borrego, S. in *A New Island Biogeography of the Sea of Cortes* (eds Case, T. J., Cody, M. L. & Ezcurra, E.) 41–59 (Oxford Univ. Press, New York, 2002).
9. Beron-Vera, F. J. & Ripa, P. Three-dimensional aspects of the seasonal heat balance in the Gulf of California. *J. Geophys. Res.* **105**, 11441–11457 (2000).
10. Galloway, J. N. *et al.* The nitrogen cascade. *Bioscience* **53**, 341–356 (2003).
11. Tilman, D. G. *et al.* Forecasting agriculturally driven global environmental change. *Science* **292**, 281–284 (2001).
12. Naylor, R., Falcon, W. P. & Puente-Gonzalez, A. *Policy Reforms and Mexican Agriculture: Views from the Yaqui Valley* (CIMMYT Economics Program Paper no. 01–01, CIMMYT, Mexico City, 2001).
13. Harrison, J. A. *Nitrogen Dynamics and Greenhouse Gas Production in Yaqui Valley Surface Drainage Waters*. PhD thesis, Stanford Univ (2003).
14. Harrison, J. A. & Matson, P. A. Patterns and controls of nitrous oxide emissions from waters draining a subtropical agricultural valley. *Glob. Biogeochem. Cycles* **17**, 1080, doi:10.1029/2002GB001991 (2003).
15. Matson, P. A., Naylor, R. & Ortiz-Monasterio, I. Integration of environmental, agronomic, and economic aspects of fertilizer management. *Science* **280**, 112–115 (1998).
16. Riley, W. J., Ortiz-Monasterio, I. & Matson, P. A. Nitrogen leaching and soil nitrate, nitrite, and ammonium levels under irrigated wheat in Northern Mexico. *Nutrient Cycling Agroecosyst.* **61**, 223–236 (2001).
17. Kahr, M., Marinone, S. G., Lluch-Cota, S. E., Pares-Sierra, A. & Mitchell, B. G. Ocean color variability in the Gulf of California: scales from days to ENSO. *Deep-Sea Res. II* **51**, 139–146 (2004).
18. Switzer, A. C., Kamykowski, D. & Zentara, S.-J. Mapping nitrate in the global ocean using remotely sensed sea surface temperature. *J. Geophys. Res.* **108**, 3280 (2003).
19. McCullagh, P. & Nelder, J. A. *Generalized Linear Models* (Chapman and Hall, London, 1989).
20. Pegau, W. S., Boss, E. & Martinez, A. Ocean color observations of eddies during the summer in the Gulf of California. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL014076 (2002).
21. Gaxiola-Castro, G., Alvarez-Borrego, S., Lavin, M., Zirino, A. & Najera-Martinez, S. Spatial variability of the photosynthetic parameters and biomass of the Gulf of California phytoplankton. *J. Plankton Res.* **21**, 231–245 (1999).
22. Mills, M. M., Ridame, C., Davey, M., La Roche, J. & Gelder, R. J. Iron and phosphorus co-limit nitrogen fixation in the eastern tropical North Atlantic. *Nature* **429**, 292–294 (2004).
23. Tyrrell, T. & Lucas, M. I. Geochemical evidence of denitrification in the Benguela upwelling system. *Continental Shelf Res.* **22**, 2497–2511 (2002).

24. Howell, E. A., Doney, S. C., Fine, R. A. & Olson, D. B. Geochemical estimates of denitrification in the Arabian Sea and the Bay of Bengal during WOCE. *Geophys. Res. Lett.* **24**, 2549–2552 (1997).
25. Valsaraj, C. P. & Rao, V. N. R. Nitrogen limitation in the tropical waters of the Bay of Bengal. *Hydrobiologia* **286**, 139–148 (1994).
26. Chen, Y. L., Chen, H.-Y., Karl, D. M. & Takahashi, M. Nitrogen modulates phytoplankton growth in spring in the South China Sea. *Continental Shelf Res.* **24**, 527–541 (2004).
27. Tilman, D. G., Cassman, K. G., Matson, P. A., Naylor, R. & Polasky, S. Agricultural sustainability and intensive production practices. *Nature* **418**, 671–677 (2002).
28. Sala, E., Aburto-Oropeza, O., Reza, M., Paredes, G. & Lopez-Lemus, L. G. Fishing down coastal food webs in the Gulf of California. *Fisheries* **29**, 19–25 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful for funding from the David and Lucile Packard Foundation and the US NSF Graduate Research Fellowship Program. Thanks to G. van Dijken and R. Labiosa for remote sensing support, I. Ortiz-Monasterio for irrigation data, E. Cruz-Colin for sharing unpublished data, and K. Beman for editorial advice. We thank P. Vitousek and R. Dunbar for comments on an earlier draft of this manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.B. (beman@stanford.edu).

The global distribution of clinical episodes of *Plasmodium falciparum* malaria

Robert W. Snow^{1,2}, Carlos A. Guerra³, Abdisalan M. Noor¹, Hla Y. Myint⁴ & Simon I. Hay^{1,3}

¹Public Health Group, KEMRI/Wellcome Trust Research Laboratories PO Box 43640, 00100 Nairobi, Kenya

²Centre for Tropical Medicine, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DS, UK

³TALA Research Group, Tinbergen Building, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

⁴Faculty of Tropical Medicine, Mahidol University, Bangkok 10400, Thailand

Interest in mapping the global distribution of malaria is motivated by a need to define populations at risk for appropriate resource allocation^{1,2} and to provide a robust framework for evaluating its global economic impact^{3,4}. Comparison of older^{5–7} and more recent^{1,4} malaria maps shows how the disease has been geographically restricted, but it remains entrenched in poor areas of the world with climates suitable for transmission. Here we provide an empirical approach to estimating the number of clinical events caused by *Plasmodium falciparum* worldwide, by using a combination of epidemiological, geographical and demographic data. We estimate that there were 515 (range 300–660) million episodes of clinical *P. falciparum* malaria in 2002. These global estimates are up to 50% higher than those reported by the World Health Organization (WHO) and 200% higher for areas outside Africa, reflecting the WHO's reliance upon passive national reporting for these countries. Without an informed understanding of the cartography of malaria risk, the global extent of clinical disease caused by *P. falciparum* will continue to be underestimated.

The Global Burden of Diseases programme of the WHO has attempted to enumerate the health consequences of malaria infection^{8,9}. Because the African region has a notoriously weak system of reporting infectious diseases, epidemiological evidence from carefully conducted prospective, 'active' case-detection studies of malaria morbidity, disability and mortality in populations living

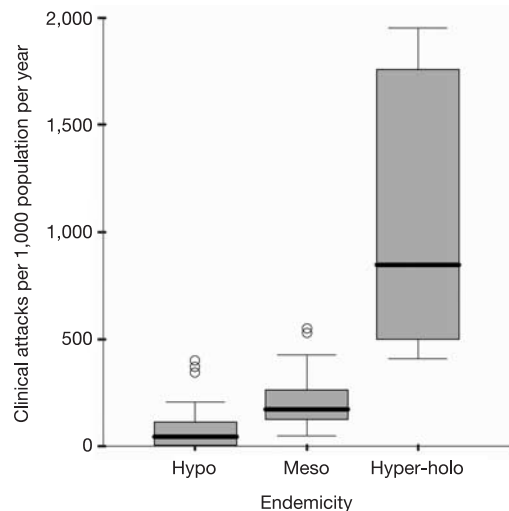


Figure 1 Annual clinical incidence of *P. falciparum* per 1,000 population according to hypoendemic ($n = 39$), mesoendemic ($n = 25$) and combined hyperendemic and holoendemic ($n = 8$) conditions. The box indicates the inter-quartile range (25% and 75%) and the thick line within the box represents the median. The whiskers represent the 2.5% and 97.5% centiles and outliers are plotted as circles outside this range. Three studies were excluded because they were undertaken in areas of a recorded zero *P. falciparum* prevalence, and each reported no clinical attacks due to *P. falciparum*.

under different transmission intensity risks have been compiled to estimate the disease burden¹⁰. A different approach was adopted for WHO regions outside Africa, where the burden was computed from 'passive' national disease and mortality notifications to WHO regional offices without precisely defining the populations exposed to varied malaria infection risks^{9,11,12}. This use of national disease registration systems to provide accurate reflections of disease rests on three assumptions: that there is complete temporal coverage (every month is reported by a facility), that there is complete spatial

coverage (every health facility reports nationwide), and that all disease events present to, and are reported by, health facilities. In reality, passive detection of disease events in most resource-poor countries is incomplete, even outside Africa.

Here we provide a standard global approach to deriving clinical malaria burden by using evidence of the epidemiological risks of disease outcome from active case-detection studies in combination with estimates of populations at risk of various *P. falciparum* transmission conditions. A comprehensive outline of these procedures is given in Methods. A conservative approach is defined to further account for the confounding of malaria diagnosis efficiency by endemicity (see Supplementary Information A for more detail and original data) and the modifying influence on endemicity of current levels of control and urbanization (see Supplementary Information B).

Our global model suggests that, in 2002, 2.2 billion people were exposed to the threat of *P. falciparum* malaria, resulting in a conservative estimate of 515 (range 300–660) million clinical attacks attributable to this parasite during that year (Fig. 1 and Table 1). At a regional level, most clinical events attributable to *P. falciparum* were concentrated in the African region (70%), but the highly populated South East Asia region contributed 25% of the world's clinical attacks in 2002 (Fig. 2 and Table 2). The WHO suggests that there were 273 million clinical attacks of malaria worldwide in 1998 and that 90% of the global disease incidence is borne by Africa⁹. Other WHO estimates report that in 1990 the global incidence of malaria was 213 million cases¹³. Neither of these sources provides sufficient detail on how the estimates were derived. Our models, by contrast, are both data-driven and reproducible. They also indicate that the number of clinical attacks due to *P. falciparum* might be 50% higher than WHO estimates, and highlight the fact that almost one-third of the global incidence occurs outside Africa.

We have not examined mortality attributed directly to *P. falciparum*, because of the paucity of prospective epidemiological descriptions of cause-specific mortality outside Africa¹⁴. The risk of death after a clinical attack of *P. falciparum* seems much higher in Africa than in South East Asia and the western Pacific. The incidence of severe, life-threatening complications of

Table 1 Populations (millions) at risk in 2002

Region	Population in <i>P. falciparum</i> endemicity classes				Total population at risk
	Unclassified	Hypoendemic	Mesoendemic	Hyperendemic and holoendemic	
Africa	13.6	39.3	67.4	414.3	521.0
Americas	3.5	43.9	10.5	0	54.5
South East Asia	47.8	827.6	486.0	0.3	1,313.9
Western Pacific	22.4	77.6	63.4	1.0	142.0
Eastern Mediterranean	32.3	143.0	33.4	0	176.4
Europe	1.1	0.3	3.2	0	3.5
Total world	120.7	1,131.7	663.9	415.6	2,211.3

Populations were determined according to classifications of malaria risk adjusted for urbanization by WHO region (see Supplementary Information B).

Table 2 Estimated data for *P. falciparum* clinical malaria cases in 2002 (millions)

Parameter	Hypoendemic	Mesoendemic	Hyperendemic and holoendemic	Total <i>P. falciparum</i> cases
Attack rate (per 1,000 population per year)	43 (6–117)	171 (125–261)	849 [500]	–
Cases per WHO region (millions)				
Africa	1.69 (0.24–4.60)	11.52 (8.42–17.58)	351.77 (207.17–351.77)	364.98 (215.82–373.95)
Americas	1.89 (0.26–5.14)	1.80 (1.32–2.75)	0	3.69 (1.58–7.89)
South East Asia	35.59 (4.97–96.83)	83.11 (60.76–126.86)	0.24 (0.14–0.24)	118.94 (65.86–223.93)
Western Pacific	3.34 (0.46–9.08)	10.84 (7.93–16.55)	0.85 (0.50–0.85)	15.03 (8.89–26.48)
Eastern Mediterranean	6.15 (0.86–16.73)	5.71 (4.17–8.71)	0	11.86 (5.03–25.44)
Europe	0.01 (0.00–0.03)	0.54 (0.40–0.83)	0	0.55 (0.40–0.86)
Total world	48.67 (6.79–132.41)	113.52 (82.99–173.28)	352.86 (207.81–352.87)	515.05 (297.59–658.55)

Results are medians and IQR (in parentheses) by WHO region and endemicity class, based on urban-adjusted denominators, derived from Table 1; the lower quartile (in square brackets) is presented instead of the IQR for populations living under conditions of hyperendemic and holoendemic transmission (see Methods).

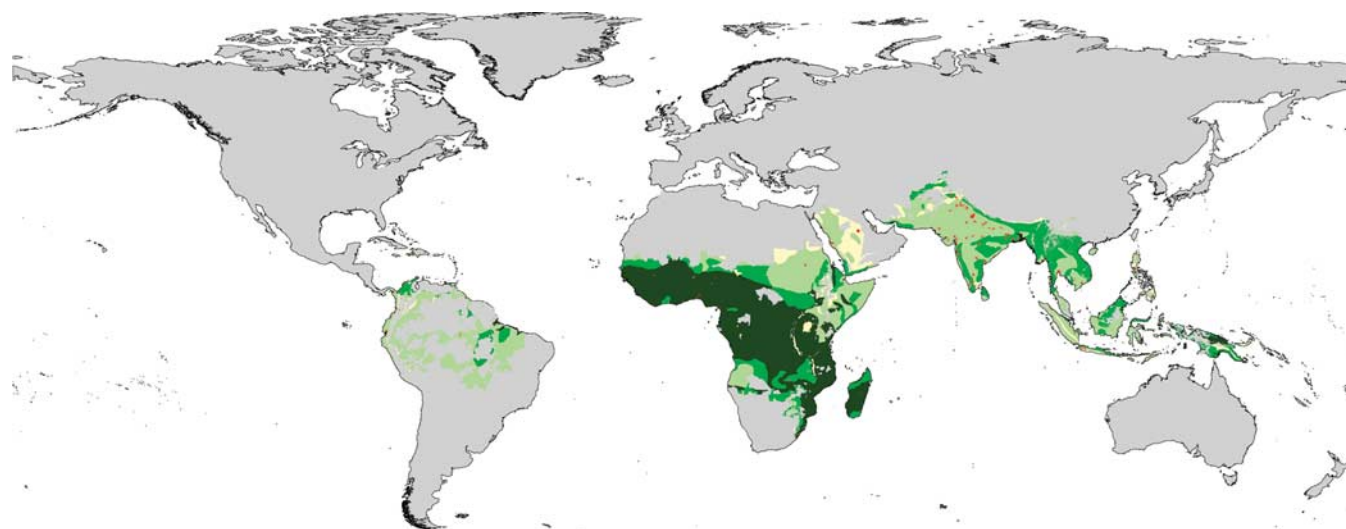


Figure 2 *P. falciparum* endemicity distribution within the global limits of risk. Endemicity classes: light green, hypoendemic (areas in which childhood infection prevalence is less than 10%); medium green, mesoendemic (areas with infection prevalence between 11% and 50%); dark green, hyperendemic and holoendemic (areas with an infection prevalence of 50% or more)¹³. Unclassified areas (yellow) represent only 6% of the global

population at risk and are due to discrepancies between the 2002 delineation of risk and the endemicity risk limits developed in refs 6 and 7. Grey areas are a combined mask of areas outside of the transmission limits and areas of population density less than 1 person km⁻² (ref. 16).

P. falciparum malaria in Africa¹⁵ is at least tenfold that in similar malaria endemic areas in India¹⁶ and Vanuatu¹⁷. Reasons for this are unclear but might include better access to prompt treatment¹⁸ and some cross-*Plasmodium* species protection against severe disease outcomes¹⁹.

We had estimated previously from epidemiological data that there were 221 million *P. falciparum* attacks in Africa in 1995 (ref. 10). Our 2002 estimate for Africa of 365 million clinical cases derives from a more specific, urban-adjusted endemicity map than that developed specifically for Africa in 1995, which was not structured according to levels of parasite prevalence. It was estimated¹² from national statistics that there were 51.2 million *P. falciparum* cases outside Africa in 1995; our estimate of 150 million cases is considerably higher. There are several possible explanations for this disparity, including our assigning populations at risk of different transmission conditions on the basis of an endemicity map constructed in 1968. We have used this map in its original form because there is no modern equivalent but have taken a very conservative approach to reclassifying areas at risk in 2002 by stepping down endemicity risks in all areas outside Africa and allowing for the rapid increases in urbanization since 1968. Furthermore, the clinical data on active detection of cases were derived from a wide range of malaria endemicities (see Supplementary Information A) to create plausible endemicity-specific median estimates of disease. It seems unlikely that we have overestimated the clinical risks when reapplied to the global distributions of the three broad endemicity classes.

The most obvious explanation is the dependence on national statistics derived from passive detection of cases for the WHO's present global disease estimates outside Africa. In our analysis we were able to compare WHO reports of clinical incidence from 12 administrative units with survey reports of data on active case detection in the same areas. These limited comparisons demonstrated the scale of under-reporting by passive detection, varying from a threefold difference in Brazil to a 1,000-fold difference in Pakistan.

The global Roll Back Malaria (RBM) initiative aims at halving the burden of malaria within the next six years⁹. The Millennium

Development Goal's target is to halt the rising incidence of malaria by 2015 (ref. 20). To achieve this, international priorities and resources must be targeted using different information sources, including national economic capacities, evidence-based cost-effective strategies and disease burdens. Inadequate descriptions of the global distribution of disease risk make it impossible to determine priorities and advise funding agencies appropriately. Redressing these deficiencies with robust data must be a priority if international agencies are to understand the size of the challenge set by their targets over the next ten years. □

Methods

To identify reports of *P. falciparum* morbidity risks defined through epidemiological studies, an electronic data search was undertaken through PubMed (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>) using the keyword 'malaria' in conjunction with each country name in all WHO regions. Abstracts were reviewed to identify reports of malaria morbidity incidence, and full papers were obtained for all original reports, or cross-referenced sources, that met the following selection criteria: first, that the report covered the period after 1985; second, that the study involved active detection of cases with the use of clinical or epidemiological morbidity definitions; third, that it was possible to compute the numbers of cases confirmed microscopically and the numbers of person-years of observation; and last, that data were reported for all age groups to capture the cumulative incidence from birth until adulthood in communities with different age-specific incidence patterns. Using these criteria for inclusion we identified 83 independent annual incidence estimates of *P. falciparum* clinical attacks from 22 countries in five WHO regions (see Supplementary Information A). No data were identified from Tajikistan, the only country in which *P. falciparum* malaria transmission occurs in the European region.

Infection prevalence has been used since the 1950s to describe malaria endemicity categorically²¹. We therefore identified coincidental cross-sectional measures of *P. falciparum* infection prevalence within the original morbidity reports, or associated publications, and matched these to 75 communities where rates of clinical attack had been established (see Supplementary Information A). *P. falciparum* annual rates of clinical attack were summarized as medians and interquartile ranges (IQR) to allow for the ranges and uncertainty of survey estimates within three infection prevalence classes: hypoendemic (parasite prevalence less than 10%), mesoendemic (parasite prevalence between 11% and 50%) and combined hyperendemic and holoendemic (parasite prevalence 50% or more) (Fig. 1). Epidemiological case definitions in areas of hyperendemicity to holoendemicity pose difficulties where fever and infection are common but are not necessarily related causally. We have adopted a working clinical definition in all malaria endemicities of fever in the presence of patent peripheral infection, accepting that this will overestimate the incidence of clinical disease in areas of hyperendemic to holoendemic transmission. To accommodate overestimation in areas of high transmission we have used the lower quartile and median as a more conservative and biologically plausible range of clinical risk.

To define the congruence of human population distribution and *P. falciparum* transmission we used spatially linked databases of human population, limits of malaria risk and malaria endemicity within a Geographic Information System (GIS) as outlined in detail in Supplementary Information B. In brief, we first defined the spatial extent of *P. falciparum* risk by using the mapped global limits of malaria risk provided by the WHO²² and modified with contemporary descriptions of spatial risk used to inform antimalarial chemoprophylaxis regimes in travellers^{22,23}, to exclude the following: countries with only *P. vivax* transmission, areas above anopheline vector-specific altitude limits, and administrative areas defined as risk-free. These boundaries formed the limits of *P. falciparum* risk and were overlaid on the only available global map of malaria endemicity developed in 1968 (refs 6, 7). This map was part of a major synthesis of historical records, documents and maps of malaria endemicity (using the hypoendemic to holoendemic classifications) interpolated globally for malaria at the peak of its assumed historical distribution. We have assumed that this endemicity map is consistent with contemporary malaria risks in Africa, but development and intervention will have substantially reduced malaria risk elsewhere¹¹. Outside the African region we consider the historical (1968) hyperendemic to holoendemic areas as contemporary (2002) mesoendemic conditions, historically mesoendemic areas as hypoendemic, and hypoendemic risk areas at their historical descriptions within the revised 2002 spatial limits of risk.

Data from Gridded Population of the World (GPW3) version 3.0 beta (<http://sedac.ciesin.colombia.edu/gpw>) were projected to 2002 by using national inter-censal growth rates from the UN Population Prospects database (<http://esa.un.org/unpp>). Population totals were extracted by country for those residing in hypoendemic, mesoendemic and hyperendemic-to-holoendemic settings (Table 1, Fig. 2). These population totals were further adjusted for the suppressive effects of urbanization on malaria transmission²⁴ by identifying all urban areas with populations of more than 1 million. Urban population totals within these pixels were reclassified to the risk class below their original classification; thus, those located in hypoendemic areas were regarded as being at no infection risk. Populations in 2002 residing in the different urban-adjusted, *P. falciparum* endemicity risk zones are shown in Table 1. The endemicity-specific morbid risks were then applied to populations within their respective endemicity classes to estimate numbers of clinical events in 2002 (Table 2).

Received 10 November; accepted 30 December 2004; doi:10.1038/nature03342.

- Hay, S. I., Guerra, C. A., Tatem, A. J., Noor, A. M. & Snow, R. W. The global distribution and population at risk of malaria: past, present and future. *Lancet Infect. Dis.* **4**, 327–336 (2004).
- Snow, R. W., Marsh, K. & le Sueur, D. The need for maps of transmission intensity to guide malaria control in Africa. *Parasitol. Today* **12**, 455–457 (1996).
- Gallup, J. L. & Sachs, J. D. The economic burden of malaria. *Am. J. Trop. Med. Hyg.* **64**, 85–96 (2001).
- Kiszewski, A. *et al.* A global index representing the stability of malaria transmission. *Am. J. Trop. Med. Hyg.* **70**, 486–498 (2004).
- Russell, P. F. World-wide malaria distribution, prevalence and control. *Am. J. Trop. Med. Hyg.* **5**, 937–956 (1956).
- Lysenko, A. J. & Semashko, I. N. in *Itogi Nauki: Medicinskaja Geografija* (ed. Lebedew, A. W.) 25–146 (Academy of Sciences, Moscow, 1968).
- Lysenko, A. J. & Beljaev, A. E. An analysis of the geographical distribution of *Plasmodium ovale*. *Bull. World Health Organ.* **40**, 383–394 (1969).
- Murray, C. J. L. & Lopez, A. D. Mortality by cause for eight regions of the world: global burden of disease study. *Lancet* **349**, 1269–1276 (1997).
- World Health Organization. *The World Health Report 1999: Making a Difference* (World Health Organization, Geneva, 1999).
- Snow, R. W., Craig, M. H., Deichmann, U. & Marsh, K. Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. *Bull. W.H.O.* **77**, 624–640 (1999).
- Carter, R. & Mendis, K. N. Evolutionary and historical aspects of the burden of malaria. *Clin. Microbiol. Rev.* **15**, 564–594 (2002).
- Mendis, K., Sina, L. J., Marchensini, P. & Carter, R. The neglected burden of *Plasmodium vivax* malaria. *Am. J. Trop. Med. Hyg.* **64**, 97–106 (2001).
- Murray, C. J. & Lopez, A. D. *Global Health Statistics: a Compendium of Incidence, Prevalence and Mortality Estimates for over 200 Countries* (Harvard School of Public Health, Boston/World Health Organization, Geneva, 1996).
- International Network for the Continuous Demographic Evaluation of Populations and Their Health in Developing Countries. (<http://www.INDEPTH-network.org>).
- Snow, R. W. *et al.* Relation between severe malaria morbidity in children and level of *Plasmodium falciparum* transmission in Africa. *Lancet* **349**, 1650–1654 (1997).
- Prusty, S. K. R. & Das, B. S. Low incidence of the severe complications of malaria and absence of malaria-specific mortality, in Tensa, Sundergarh district, Orissa State, India, an area hyper-endemic for malaria. *Ann. Trop. Med. Parasit.* **95**, 133–140 (2001).
- Maitland, K. *et al.* Absence of malaria-specific mortality in children in an area of hyperendemic malaria. *Trans. R. Soc. Trop. Med. Hyg.* **91**, 562–566 (1997).
- Alles, H. K., Mendis, K. N. & Carter, R. Malaria mortality rates in South Asia and in Africa: implications for malaria control. *Parasitol. Today* **14**, 369–375 (1998).
- Mayfong, M., Pukrittayakamee, S., Newton, P. N. & White, N. J. Mixed species malaria infections in humans. *Trends Parasitol.* **20**, 233–240 (2004).
- United Nations Development Programme. *Human Development Report 2003. Millennium Development Goals: a Compact among Nations to End Human Poverty* (Oxford Univ. Press, Oxford, 2003).
- Metselaar, D. & Van Thiel, P. H. Classification of malaria. *Trop. Geogr. Med.* **11**, 157–161 (1959).
- World Health Organization. *International Travel and Health. Situation as on 1 January 2003* (World Health Organization, Geneva, 2003).
- International Association for Medical Assistance to Travellers. World malaria risk chart (including geographical distribution of principal vectors, geographical distribution of *P. falciparum* malaria, areas where *Plasmodium falciparum* is resistant to chloroquine and guidelines for suppressive medication by country). Status as of 15 March 2004. (<http://www.iamat.org>).

- Hay, S. I., Guerra, C. A., Tatem, A. J., Atkinson, P. M. & Snow, R. W. Urbanization, malaria transmission and disease burden in Africa. *Nature Rev. Microbiol.* **3**, 81–90 (2005).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This research was funded by the Wellcome Trust, UK. Part of the work on mapping population risk distributions was funded by the World Health Organization Roll Back Malaria Department. The World Health Organization/RBM Department Monitoring and Evaluation Team and staff from WHO regional offices provided passive case detection data, the WHO Evidence for Information and Policy Department provided SALB data and the WHO Public Health Mapping Group provided other geographical boundary files. We thank K. Marsh, D. Forster, C. Macintosh, K. Maitland and D. Rogers for comments on earlier drafts of the manuscript, D. Balk and G. Yetman for supplying alpha versions of GPW 3.0, and A. Tatem for the urban extractions. R.W.S. is a Wellcome Trust Senior Research Fellow and acknowledges the support of the Kenyan Medical Research Institute (KEMRI). S.I.H. is funded by a Research Career Development Fellowship from the Wellcome Trust. C.A.G. is partially funded by the Fundación para la Ciencia y Tecnología (FUNDACYT).

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com.

Correspondence and requests for materials should be addressed to R.W.S. (rsnow@wtinairobi.mimcom.net).

Mediation of pathogen resistance by exudation of antimicrobials from roots

Harsh P. Bais¹, Balakrishnan Prithiviraj¹, Ajay K. Jha¹, Frederick M. Ausubel² & Jorge M. Vivanco¹

¹Department of Horticulture and Landscape Architecture, and Center for Rhizosphere Biology, Colorado State University, Fort Collins, Colorado 80523-1173, USA

²Department of Genetics, Harvard Medical School, and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

Most plant species are resistant to most potential pathogens. It is not known why most plant–microbe interactions do not lead to disease, although recent work indicates that this basic disease resistance is multi-factorial^{1,2}. Here we show that the exudation of root-derived antimicrobial metabolites by *Arabidopsis thaliana* confers tissue-specific resistance to a wide range of bacterial pathogens. However, a *Pseudomonas syringae* strain that is both at least partly resistant to these compounds and capable of blocking their synthesis/exudation is able to infect the roots and cause disease. We also show that the ability of this *P. syringae* strain to block antimicrobial exudation is dependent on the type III secretory system.

Recent work has shown that the Gram-negative bacterial pathogen *P. syringae* pv. *tomato* strain DC3000 (*Pst* DC3000) infects and colonizes *A. thaliana* roots, causing extensive necrosis and ultimately killing the plants³. In contrast, seven other *P. syringae* strains that were tested (*P. syringae* pv. *phaseolicola* strains NPS3121 and race 6 (*Psp* NPS3121 and *Psp* rc6), *P. syringae* pv. *glycinea* strain A29-2 race 4 (*Psg* A29-2), *P. syringae* pv. *syringae* strain B728a (*Pss* B728a) and *P. syringae* pv. *maculicola* strains ES4326, M₁ and M₄ (*Psm* ES4326, *Psm* M₁ and *Psm* M₄)) did not cause significant root necrosis or plant mortality when inoculated into liquid medium (Fig. 1a and Supplementary Fig. S1) or sterilized planting mix (Figs 1b and 2a, and Supplementary Figs S2A and S3A) in which *Arabidopsis* seedlings were growing. The seven non-pathogenic *P. syringae* strains also colonized *Arabidopsis* roots relatively poorly compared with *Pst* DC3000 (Fig. 2b and Supplementary Figs S3B and S4). One of the non-pathogenic strains, *Psp* NPS3121, has previously been described as an *Arabidopsis* ‘non-host’ pathogen

because it does not cause disease in *Arabidopsis* leaves⁴. However, even though *Psm* ES4326 and *Psm* M₄ failed to cause disease in either of the root infection assays described here, they are just as virulent as *Pst* DC3000 when infiltrated into *Arabidopsis* leaves^{3–7}.

Our previous studies showed that *Arabidopsis* roots exude a variety of antimicrobial compounds^{8,9}. To test whether these compounds confer resistance to the seven non-pathogenic *P. syringae* strains, we repeated the root pathogenicity assays in the presence of activated charcoal, which adsorbs root-exuded secondary compounds^{10,11}. Figures 1c and 2a, b (and Supplementary Figs S2A and S3A, B) show that disease symptoms, plant mortality and root colonization increased markedly in plants inoculated with the seven non-pathogenic *P. syringae* strains in the presence of activated charcoal. Activated charcoal in the absence of the bacterial strains was not toxic to *Arabidopsis* (Fig. 1c and Supplementary Fig. S2A). Washing the activated charcoal off the roots restored resistance to two non-pathogenic strains tested (*Psp* NPS3121 and *Psg* A29-2; Supplementary Fig. S2B). Confocal scanning laser microscopy analysis showed that *Psp* NPS3121 and *Psg* A29-2 efficiently colonized activated charcoal-treated roots (Supplementary Fig. S4). In contrast, *Sinorhizobium meliloti*, a well-characterized non-pathogenic soil bacterium, did not elicit any disease symptoms in the presence of activated charcoal (Fig. 1b, c). These results support the hypothesis that root exudates have a function in rhizosphere-mediated resistance to specific bacterial pathogens.

We used high-performance liquid chromatography (HPLC) to profile low-molecular-mass compounds in root exudates of *Arabidopsis* seedlings that had been infected *in vitro*, to determine whether there is a correlation between the exudation of antimicrobial compounds and resistance to particular *P. syringae* strains. The kinetics of accumulation of a representative antimicrobial compound (butanoic acid) in root exudates from plants infected with *Pst* DC3000 (pathogenic) or *Psp* NPS3121 (representative non-pathogenic strain) compared with non-infected plants is shown in Fig. 3a. The kinetics of accumulation of nine additional antimicrobial compounds (see Table 1 and Supplementary Table S1) as well as the entire chromatographic profiles of root exudates after infection with a variety of *P. syringae* strains are shown in

Supplementary Figs S5–S7. The exudates from plants infected with the non-pathogenic strains *Psp* NPS3121 and *Psg* A29-2 contained more low-molecular-mass compounds than exudates from non-infected plants. At later time points the exudates from plants infected with *Pst* DC3000 contained even lower concentrations of compounds than exudates from non-infected plants (Fig. 3a and Supplementary Figs S5 and S6). Because the same compounds are exuded by *Arabidopsis* roots in both the presence and absence of bacterial infection, it is very unlikely that they are synthesized by the *P. syringae* strains. Much higher concentrations of some of these compounds (up to 800-fold higher for *p*-coumaric acid) were found in the root exudates of plants inoculated with *Psp* N3121 and *Psg* A29-2 than in the exudates of plants inoculated with *Pst* DC3000 (Supplementary Figs S5–S7; see Supplementary Fig. S7F).

We compared the bacteriostatic activity of total root exudates isolated from non-infected plants to exudates isolated from plants infected with *Pst* DC3000, *Psp* NPS3121 or *Psg* A29-2. We also tested the growth inhibitory activity of a cocktail containing a mixture of the ten identified compounds (Supplementary Table S1), as well as the activity of each of the ten individual compounds. The root exudates from non-infected plants had moderate bacteriostatic activity against all seven of the non-pathogenic *P. syringae* strains but did not show growth inhibition against *Pst* DC3000 (Table 1 and Supplementary Fig. S8). The root exudates elicited by *Psp* NPS3121 and *Psg* A29-2 had more bacteriostatic activity against the seven non-pathogenic strains than the exudates from non-infected plants (Table 1 and Supplementary Fig. S8). *Pst* DC3000 had a significant amount of resistance to these latter exudates (Table 1 and Supplementary Fig. S8). Similarly, the cocktail of the ten identified compounds also had bacteriostatic activity against all of the non-infectious strains but not against *Pst* DC3000 (Table 1 and Supplementary Fig. S8). When tested individually, each of the ten identified compounds, with the exception of ferulic acid, had a differential bacteriostatic activity against the seven non-infectious *P. syringae* strains (but not against *Pst* DC3000) at a concentration similar to that found in the exudates of control plants not exposed to microbes (Table 1 and Supplementary Fig. S8). The root exudates from *Arabidopsis* plants infected with *Pst* DC3000 did not contain

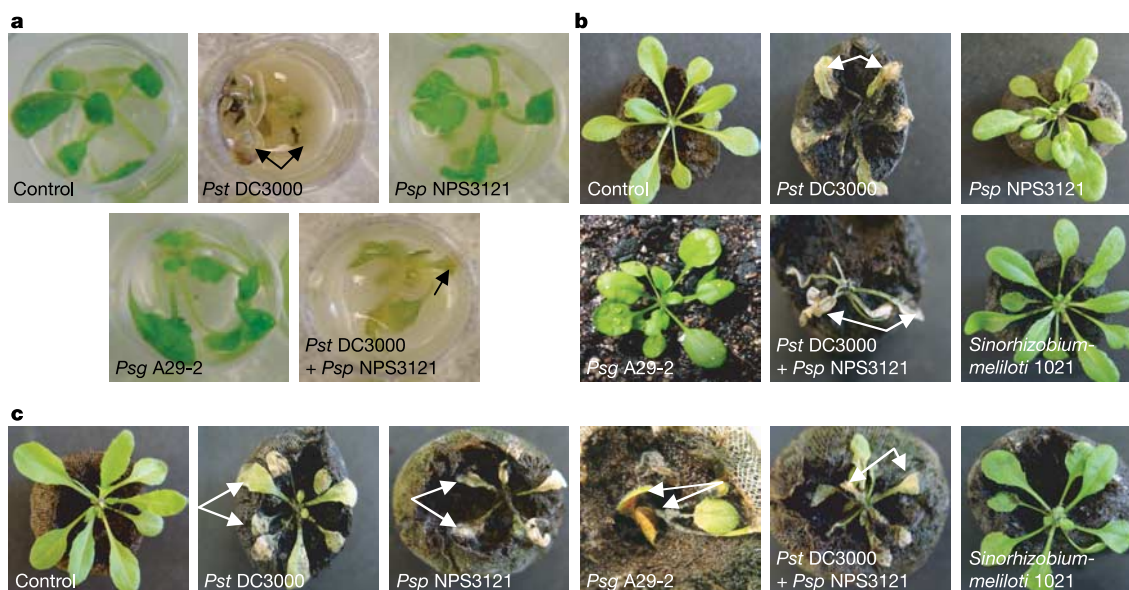


Figure 1 Pathogenicity of various *P. syringae* pathovars against *A. thaliana*. *Arabidopsis* plants grown in liquid medium (**a**), sterilized planting mix (**b**) or sterilized planting mix supplemented with activated charcoal (30 mg per gram of planting mix) (**c**) were infected with the indicated *P. syringae* strains. The appearances of disease symptoms on roots and

leaves infected with specific *P. syringae* strains are described in Supplementary Figs S1 and S2. In all cases, disease symptoms were photographed on the seventh day after inoculation. Experiments were repeated twice, each experiment containing three replicates of 20 plants each. Arrows indicate chlorosis and plant mortality.

significant bacteriostatic activity against any of the *P. syringae* strains used in this study (Table 1 and Supplementary Fig. S8).

These data suggest two potentially additive reasons why *Pst* DC3000 is pathogenic in the root infection assays. First, very low levels of antimicrobial compounds are exuded after infection with *Pst* DC3000, possibly because DC3000 escapes detection by the plant, blocks the synthesis or exudation of antimicrobials or actively degrades them. Second, *Pst* DC3000 is relatively resistant to the antimicrobial compounds, possibly because of inherent resistance, active detoxification or potent efflux systems¹².

It seems unlikely that *Pst* DC3000 simply kills the root cells before they can exude compounds. As shown in Fig. 3 and Supplementary Figs S5–S7, significant amounts of the antimicrobial compounds accumulate by 3 days after infection with the non-pathogenic strain *Psp* NPS3121, at about the time that necrotic lesions first appear on the roots of plants infected with *Pst* DC3000.

We performed co-inoculation experiments with equal mixtures of *Pst* DC3000 and *Psp* NPS3121 or *Pst* DC3000 and *Psg* A29-2 to determine whether *Pst* DC3000 actively blocks the synthesis or exudation of antimicrobial compounds. In both cases the mixed inoculum resulted in disease symptoms (Fig. 1a, b and

Supplementary Figs S1 and S2A), plant mortality (Fig. 2a and Supplementary Fig. S3A), proliferation of both of the inoculating *P. syringae* strains in the rhizosphere (Fig. 2b and Supplementary Fig. S3B; data not shown), and a lack of antimicrobial compounds in root exudates (Fig. 3a and Supplementary Figs S6 and S7). These data suggest that *Pst* DC3000 either actively blocks the synthesis or exudation of antimicrobials or actively degrades these compounds.

When root exudates or the ten individual compounds were tested for their ability to inhibit the growth of a mixture of *Pst* DC3000 and *Psp* NPS3121 or *Pst* DC3000 and *Psg* A29-2, strains *Psp* NPS3121 or *Psg* A29-2 were inhibited to levels that were similar to those when tested without *Pst* DC3000, whereas *Pst* DC3000 was able to grow (see Supplementary Methods; data not shown). This makes it unlikely that *Pst* DC3000 is actively causing the degradation of antimicrobial compounds in the exudates.

Additional experiments were performed to determine whether the *Pst* DC3000 type III secretion system (TTSS), which involves the translocation of bacterial-encoded virulence factors directly into host cells, has a function in blocking the synthesis or exudation of antimicrobial compounds. The TTSS is encoded by *hrp* (for hypersensitive response and pathogenesis) and *hrc* (for hypersensitive response conserved) genes¹³ and is required for pathogenicity in a variety of infection assays¹⁴. As shown in Fig. 3b and Supplementary Figs S5 and S6, *Pst* DC3000 *hrcC* and *hrpL* mutants elicited profiles of compounds similar to those generated by *Psp* NPS3121 and *Psg* A29-2 inoculation. The HrcC protein is a structural component of the TTSS machinery. The HrpL protein is required for the transcription of *hrcC* and of most of the *hrp* genes as well as a variety of type III effectors^{15,16} and virulence genes, including genes

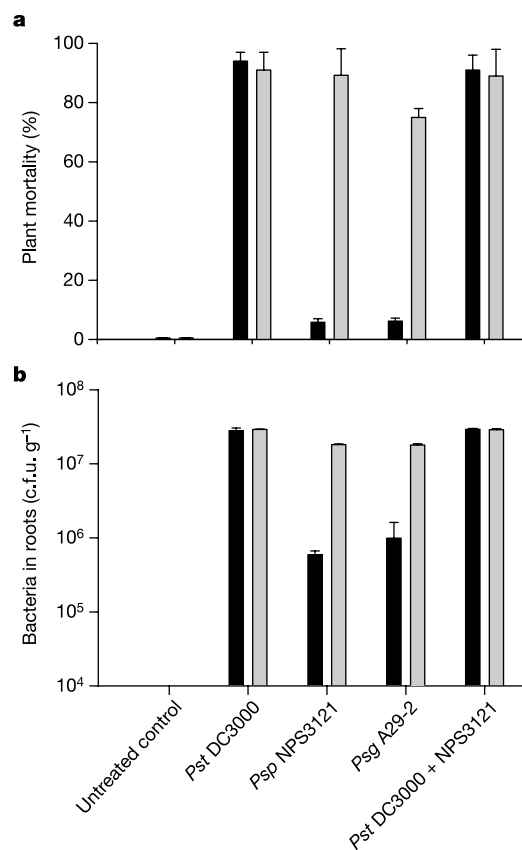


Figure 2 Activated charcoal enhances virulence of non-host pathogens. **a**, Plant mortality at 7 days after infection. Two-way analysis of variance for plant mortality: $F_{\text{treatment}}=14.12$; $df = 1,56$; $P < 0.005$. **b**, Growth of *P. syringae* in the rhizosphere. Data are averages of two independent experiments. Experiments were conducted in the presence (grey bars) and absence (black bars) of activated charcoal. In panel **b** for the mixed inoculation treatment (*Pst* DC3000 + *Psp* NPS3121), the group bars depict c.f.u. numbers for *Pst* DC3000. The c.f.u. numbers for *Psp* NPS3121 in untreated and activated charcoal treatment under the mixed inoculation regime were about 3.9×10^5 and about 3.1×10^5 bacteria per gram of roots respectively. Error bars indicate s.e.m. Two-way analysis of variance for total bacterial counts: $F_{\text{treatment}}=12.56$; $df = 1,38$; $P < 0.001$.

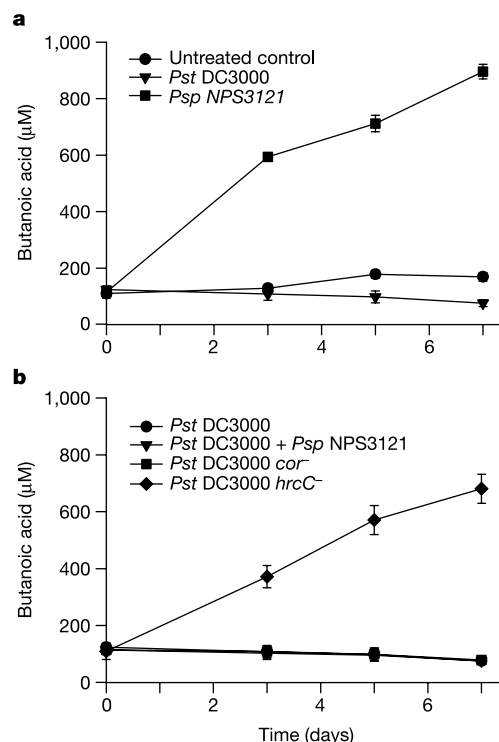


Figure 3 Kinetics of accumulation of a representative antimicrobial compound (butanoic acid) in *Arabidopsis* root exudates. **a**, **b**, *Arabidopsis* plants infected with *Pst* DC3000, *Psp* NPS3121 (representative non-pathogenic strain), a mixture of *Pst* DC3000 and *Psp* NPS3121, *Pst* DC3000 *cor*⁻ or *Pst* DC3000 *hrcC*⁻ are compared with non-infected plants on days 0, 3, 5 and 7 after inoculation. Data are the average of ten independent experiments. Error bars indicate s.e.m. Two-way analysis of variance for each metabolite concentration: $F_{\text{treatment}}=12.41$; $df = 1,58$; $P < 0.001$.

Table 1 Bacteriostatic activities of total root exudates and secondary metabolites isolated from *Arabidopsis* root exudates

Compounds	Bacterial growth inhibition (%)			
	<i>Pst</i> DC3000	<i>Psp</i> NPS3121	<i>Psg</i> A29-2	<i>Pst</i> DC3000 <i>hrcC</i> [−]
Total root exudates (Untreated)	1.52 ± 0.07	84.51 ± 2.52	74.61 ± 4.83	19.22 ± 1.24
Total root exudates (<i>Pst</i> DC3000)	1.54 ± 0.09	2.13 ± 0.09	1.82 ± 0.02	1.41 ± 0.022
Total root exudates (<i>Psp</i> NS3121)	7.13 ± 0.15	91.24 ± 2.81	81.91 ± 3.52	36.91 ± 3.91
Total root exudates (<i>Psg</i> A29-2)	8.91 ± 0.21	95.31 ± 4.12	83.21 ± 4.83	41.82 ± 2.14
Cocktail of compounds listed below	0.82 ± 0.03	78.93 ± 8.94	69.12 ± 7.14	34.91 ± 4.04
1. Butanoic acid	0.74 ± 0.02	46.91 ± 8.02	48.14 ± 12.13	17.12 ± 3.91
2. <i>trans</i> -Cinnamic acid	0.41 ± 0.01	44.21 ± 8.24	22.21 ± 9.13	31.64 ± 4.23
3. <i>o</i> -Coumaric acid	0.33 ± 0.05	51.21 ± 7.14	41.31 ± 3.81	9.83 ± 1.94
4. <i>p</i> -Coumaric acid	0.12 ± 0.03	18.12 ± 4.03	18.24 ± 2.14	16.91 ± 3.82
5. 3-Indolepropanoic acid	0.63 ± 0.04	89.14 ± 11.01	4.13 ± 0.02	28.84 ± 5.21
6. Ferulic acid	0.14 ± 0.04	0	0	14.81 ± 4.12
7. Vanillic acid	0.21 ± 0.04	1.23 ± 0.04	46.61 ± 5.12	11.53 ± 1.52
8. Methyl <i>p</i> -hydroxybenzoate	0.94 ± 0.04	21.23 ± 6.12	7.11 ± 1.14	18.91 ± 2.90
9. <i>p</i> -Hydroxybenzamide	0.73 ± 0.07	91.14 ± 3.41	6.41 ± 1.22	21.84 ± 1.93
10. Syringic acid	1.71 ± 0.08	19.21 ± 6.32	3.21 ± 0.03	19.82 ± 1.82

Net bacterial growth was calculated by subtracting the final OD₆₀₀ from the initial OD₆₀₀ readings after 24 h of incubation. Percentage inhibition was calculated from net bacterial growth based on OD₆₀₀ readings by the following formula: $100 \times [(\text{untreated} - \text{treated}) / \text{untreated}]$. Results are means $\pm$ s.e.m. for five independent experiments with five replicates each. Two-way analysis of variance for percentage inhibition: $F_{\text{treatment}} = 13.23$; $df = 1, 41$; $P < 0.005$. The concentration of antimicrobial compounds in all the treatments approximates the biological concentration of the same compounds (see Supplementary Figs S5–S7) in uninfected *Arabidopsis* root exudates. A complementary graph with all the non-pathogenic *P. syringae* strains is provided in Fig. 3 and Supplementary Figs S5–S7. The compounds numbered 1–10 here are referred to by these numbers in Supplementary Table S1 and Supplementary Figs S5–S7.

involved in the biosynthesis of the phytotoxin coronatine¹⁷. However, a *Pst* DC3000 mutant deficient in coronatine production elicited disease symptoms (Supplementary Fig. S1) and suppressed the antimicrobial compounds in the root exudates similarly to *Pst* DC3000 (Fig. 3b, Supplementary Figs S5–S7). Moreover, coronatine at concentrations as high as 20 $\mu\text{g ml}^{-1}$ in the liquid medium failed to suppress—and in fact stimulated—the exudation of antimicrobial compounds (data not shown). These results suggest that the TTSS and perhaps other virulence factors under HrpL control (but not coronatine) are required for blocking the synthesis or exudation of antimicrobial compounds.

The data in Table 1 and Supplementary Fig. S8 show that the *hrcC* and *hrpL* mutants are more susceptible than their *Pst* DC3000 parent to the root exudates from non-infected plants as well as to the exudates elicited by *Psp* NPS3121 and *Psg* 29-2. The *hrcC* and *hrpL* mutants were also more susceptible to the cocktail of ten isolated compounds and the individually administered compounds (Table 1, Supplementary Fig. S8). These data suggest that the TTSS in *P. syringae* is also important in conferring resistance to plant-derived antibiotics.

Many studies have described the existence of preformed and induced low-molecular-mass compounds referred to as phyto-anticipins or phytoalexins, respectively, that have antimicrobial activity^{18–22}. Here we have shown that in addition to constitutive active defences, infection of *Arabidopsis* with a variety of *P. syringae* strains elicits the exudation of significantly higher concentrations of these antimicrobials. The consequence of both the constitutive and inducible production of antimicrobials is that *Arabidopsis* roots are resistant to seven of eight diverse *P. syringae* strains tested. However, the eighth strain, *Pst* DC3000, is a potent *Arabidopsis* root pathogen under our experimental conditions. In *Pst* DC3000, the TTSS not only seems to block the exudation of antimicrobial compounds but also confers at least partial resistance to these compounds. How this occurs awaits further explanation. A note of caution is that *Pst* DC3000 is not generally considered a genuine root pathogen and that further studies are required to determine whether our laboratory observations are relevant to natural habitats where many microbes compete in the rhizosphere. Recently, in results reminiscent of our study, *Magnaporthe grisea*, a fungal leaf pathogen of rice, was also shown to be capable of infecting rice roots²³.

It therefore seems that *Pst* DC3000 has a fail-safe mechanism for overcoming host defences. We suggest that because *Pst* DC3000 is inherently more resistant to exuded antimicrobial compounds, it is

able to proliferate in the rhizosphere and to initiate a root infection. This enables *Pst* DC3000 in a second step to block the further production of antimicrobials by the translocation of type III effectors into root cells. These findings might offer a general explanation for the observation that very few bacterial pathogens are able to cause disease on any particular host plant. □

Methods

Plant and bacterial strains, and growth conditions

Wild-type *Arabidopsis thaliana* ecotype Columbia (Col-0) seeds (Lehle Seeds) were surface-sterilized with sodium hypochlorite (0.3% v/v) and germinated on solidified Murashige and Skoog basal medium (MS)²⁴ in a growth chamber at 25 °C. *P. syringae* pv. *tomato* DC3000 (*Pst* DC3000)²⁵, *P. syringae* pv. *phaseolicola* NPS3121 (*Psp* NPS3121)²⁶, *P. syringae* pv. *glycinea* 29-2 race 4 (*Psg* A29-2)²⁵, *P. syringae* pv. *phaseolicola* race 6 (*Psp* rc6)²⁷, *P. syringae* pv. *syringae* B728a (*Pss* B728a)²⁸ and *P. syringae* pv. *maculicola* strains ES4326 (*Psm* ES4326)⁵, M₁ (*Psm* M₁)⁶ and M₄ (*Psm* M₄)⁷ have been described previously. *Pst* DC3000 *hrpL*[−] and *Pst* DC3000 *hrcC*[−] (ref. 16) mutants were obtained from S.-Y. He. Purified coronatine and a coronatine non-producing mutant of *Pst* DC3000 with Tn5 inserted in *cja-7* (ref. 29) (strain DB5A6) were obtained from C. Bender. *Sinorhizobium meliloti* 1021 strain was obtained from A. M. Hirsch and was cultured in Luria–Bertani (LB) medium.

Root pathogenicity assays *in vitro* and in planting mix

Arabidopsis seedlings 25 days old, growing in MS medium in 12-well tissue culture plates (see above), were inoculated with *P. syringae* strains that had been grown to an OD₆₀₀ of 0.02–0.04 in LB medium at 28 °C to give an initial dose of about 2.5×10^7 colony-forming units (c.f.u.) ml^{-1} . Twenty plants per treatment were used for the analysis of disease symptoms and mortality rates. Pots of sterilized planting mix, each containing a single 4-week-old *Arabidopsis* plant, were flooded with 10 ml of a *P. syringae* or *S. meliloti* culture grown to an OD₆₀₀ of 0.2–0.4 in LB medium at 28 °C to give an inoculum of about 10^8 to 5×10^8 c.f.u. per gram of planting mix. Twenty plants per treatment were used for the analysis of mortality rates.

In some experiments, activated charcoal (30 mg per gram of planting mix; Merck) was added to sterilized planting mix by mixing it with water (1.5 g in 5 ml of water for 50 g of planting mix) and pipetting the suspension around 4-week-old *Arabidopsis* rosettes.

Media extraction and HPLC analysis

Liquid medium samples from *Arabidopsis* plants grown *in vitro* with well-differentiated roots were freeze-dried (Genesis 25LL; Virtis), extracted and analysed by HPLC as described previously⁹.

Growth inhibition assays

Broth microdilution antimicrobial susceptibility testing was performed in accordance with published methods of the National Committee for Clinical Laboratory Standards³⁰.

Other protocols

Detailed protocols for growing *Arabidopsis* in liquid MS medium or in planting mix, for performing the root pathogenicity assays in liquid or planting soil, for determining the bacterial colonization of roots, for confocal microscopy, for activated charcoal supplementation of planting mix, for HPLC analysis and for *P. syringae* growth inhibition assays are provided in Supplementary Methods.

Received 13 May 2004; accepted 10 January 2005; doi:10.1038/nature03356.

- Nishimura, M. & Somerville, S. Plant biology. Resisting attack. *Science* **295**, 2032–2033 (2002).
- Thordal-Christensen, H. Fresh insights into processes of nonhost resistance. *Curr. Opin. Plant Biol.* **6**, 351–357 (2003).
- Bais, H. P., Fall, R. & Vivanco, J. M. Biocontrol of *Bacillus subtilis* against infection of *Arabidopsis* roots by *Pseudomonas syringae* is facilitated by biofilm formation and surfactin production. *Plant Physiol.* **134**, 307–319 (2004).
- Yu, G. L., Katagiri, F. & Ausubel, F. M. *Arabidopsis* mutations at the *RPS2* locus result in loss of resistance to *Pseudomonas syringae* strains expressing the avirulence gene *avrRpt2*. *Mol. Plant Microbe Interact.* **6**, 434–443 (1993).
- Davis, K. R., Schott, E. & Ausubel, F. M. Virulence of selected phytopathogenic pseudomonads in *Arabidopsis thaliana*. *Mol. Plant Microbe Interact.* **4**, 477–488 (1991).
- Dong, X., Mindrinos, M., Davis, K. R. & Ausubel, F. M. Induction of *Arabidopsis* defense genes by virulent and avirulent *Pseudomonas syringae* strains and by a cloned avirulence gene. *Plant Cell* **3**, 61–72 (1991).
- Debener, T., Lehnackers, H., Arnold, M. & Dangel, J. L. Identification and molecular mapping of a single *Arabidopsis thaliana* locus determining resistance to a phytopathogenic *Pseudomonas syringae* isolate. *Plant J.* **1**, 289–302 (1991).
- Walker, T. S., Bais, H. P., Grotewold, E. & Vivanco, J. M. Root exudation and rhizosphere biology. *Plant Physiol.* **132**, 44–51 (2003).
- Walker, T. S., Bais, H. P., Halligan, K. M., Stermitz, F. R. & Vivanco, J. M. Metabolic profiling of root exudates of *Arabidopsis thaliana*. *J. Agric. Food Chem.* **51**, 2548–2554 (2003).
- Callaway, R. M. & Aschehoug, E. T. Invasive plants versus their new and old neighbors: a mechanism for exotic invasion. *Science* **290**, 521–523 (2000).
- Bais, H. P., Vepachedu, R., Gilroy, S., Callaway, R. M. & Vivanco, J. M. Allelopathy and exotic plant invasion: from molecules and genes to species interactions. *Science* **301**, 1377–1380 (2003).
- Nikaido, H. Multiple antibiotic resistance and efflux. *Curr. Opin. Microbiol.* **1**, 516–523 (1998).
- Galan, J. E. & Collmer, A. Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* **284**, 1322–1328 (1999).
- Roine, E. *et al.* Hrp pilus: an hrp-dependent bacterial surface appendage produced by *Pseudomonas syringae* pv. tomato DC3000. *Proc. Natl Acad. Sci. USA* **94**, 3459–3464 (1997).
- Xiao, Y., Heu, S., Yi, J., Lu, Y. & Hutcheson, S. W. Identification of a putative alternate sigma factor and characterization of a multicomponent regulatory cascade controlling the expression of *Pseudomonas syringae* pv. *syringae* Pss61 hrp and hrpA genes. *J. Bacteriol.* **176**, 1025–1036 (1994).
- Zwiesler-Vollick, J. *et al.* Identification of novel hrp-regulated genes through functional genomic analysis of the *Pseudomonas syringae* pv. tomato DC3000 genome. *Mol. Microbiol.* **45**, 1207–1218 (2002).
- Bender, C. L., Alarcon-Chaidez, F. & Gross, D. C. *Pseudomonas syringae* phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. *Microbiol. Mol. Biol. Rev.* **63**, 266–292 (1999).
- Dixon, R. A. Natural products and plant disease resistance. *Nature* **411**, 843–847 (2001).
- Papadopoulos, K., Melton, R. E., Leggett, M., Daniels, M. J. & Osbourn, A. E. Compromised disease resistance in saponin-deficient plants. *Proc. Natl Acad. Sci. USA* **96**, 12923–12928 (1999).
- Haralampidis, K. *et al.* A new class of oxidosqualene cyclases directs synthesis of antimicrobial phytoprotectants in monocots. *Proc. Natl Acad. Sci. USA* **98**, 13431–13436 (2001).
- Reimann, C. & VanEtten, H. D. Cloning and characterization of the PDA6–1 gene encoding a fungal cytochrome P-450 which detoxifies the phytoalexin pisatin from garden pea. *Gene* **146**, 221–226 (1994).
- Bouarab, K., Melton, R., Peart, J., Baulcombe, D. & Osbourn, A. A saponin-detoxifying enzyme mediates suppression of plant defences. *Nature* **418**, 889–892 (2002).
- Sesma, A. & Osbourn, A. E. The rice leaf blast pathogen undergoes developmental processes typical of root-infecting fungi. *Nature* **431**, 582–586 (2004).
- Murashige, T. & Skoog, F. A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol. Plant.* **15**, 473–497 (1962).
- Whalen, M. C., Innes, R. W., Bent, A. F. & Staskawicz, B. J. Identification of *Pseudomonas syringae* pathogens of *Arabidopsis* and a bacterial locus determining avirulence on both *Arabidopsis* and soybean. *Plant Cell* **3**, 49–59 (1991).
- Lindgren, P. B., Peet, R. C. & Panopoulos, N. J. Gene cluster of *Pseudomonas syringae* pv. 'phaseolicola' controls pathogenicity of bean plants and hypersensitivity of nonhost plants. *J. Bacteriol.* **168**, 512–522 (1986).
- Fillingham, A. J. *et al.* Avirulence genes from *Pseudomonas syringae* pathovars *phaseolicola* and *pisi* confer specificity towards both host and non-host species. *Physiol. Mol. Plant Pathol.* **40**, 1–15 (1992).
- Rich, J. J., Hirano, S. S. & Willis, D. K. Pathovar-specific requirement for the *Pseudomonas syringae* *lrmA* gene in disease lesion formation. *Appl. Environ. Microbiol.* **58**, 1440–1446 (1992).
- Brooks, D. M. *et al.* Identification and characterization of a well-defined series of coronatine biosynthetic mutants of *Pseudomonas syringae* pv. tomato DC3000. *Mol. Plant Microbe Interact.* **17**, 162–174 (2004).
- National Committee for Clinical Laboratory Standards (NCCLS). *Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria that Grow Aerobically. Approved Standard M7–A6 5th edn* (NCCLS, Wayne, Pennsylvania, USA, 2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank W. Songnuan and J. Dangel for critical reading of the manuscript, and E. Wortman-Wunder for editing suggestions. This work was supported by the Colorado State University Agricultural Experiment Station (J.M.V.) and by NIH and NSF grants to F.M.A. J.M.V. is a NSF-CAREER Faculty Fellow.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.V. (j.vivanco@colostate.edu) or F.M.A. (ausubel@molbio.mgh.harvard.edu).

Spike-timing-dependent synaptic plasticity depends on dendritic location

Robert C. Froemke, Mu-ming Poo & Yang Dan

Department of Molecular and Cell Biology and Helen Wills Neuroscience Institute, University of California, Berkeley, California 94720-3200, USA

In the neocortex, each neuron receives thousands of synaptic inputs distributed across an extensive dendritic tree. Although postsynaptic processing of each input is known to depend on its dendritic location^{1–8}, it is unclear whether activity-dependent synaptic modification is also location-dependent. Here we report that both the magnitude and the temporal specificity of spike-timing-dependent synaptic modification^{9–17} vary along the apical dendrite of rat cortical layer 2/3 pyramidal neurons. At the distal dendrite, the magnitude of long-term potentiation is smaller, and the window of pre-/postsynaptic spike interval for long-term depression (LTD) is broader. The spike-timing window for LTD correlates with the window of action potential-induced suppression of NMDA (N-methyl-D-aspartate) receptors; this correlation applies to both their dendritic location-dependence and pharmacological properties. Presynaptic stimulation with partial blockade of NMDA receptors induced LTD and occluded further induction of spike-timing-dependent LTD, suggesting that NMDA receptor suppression underlies LTD induction. Computer simulation studies showed that the dendritic inhomogeneity of spike-timing-dependent synaptic modification leads to differential input selection at distal and proximal dendrites according to the temporal characteristics of presynaptic spike trains. Such location-dependent tuning of inputs, together with the dendritic heterogeneity of postsynaptic processing, could enhance the computational capacity of cortical pyramidal neurons.

Whole-cell recordings were made from layer 2/3 pyramidal neurons in rat cortical slices¹⁴. To activate synapses selectively at each dendritic location, two extracellular stimulating electrodes were placed <10 μm away from the apical dendrite, one <50 μm from the soma ('proximal') and one >100 μm from the soma ('distal', Fig. 1a). Both electrodes reliably evoked excitatory postsynaptic potentials (EPSPs) from separate populations of synapses, as indicated by linear EPSP summation^{3,7} (Fig. 1b) and the absence of paired-pulse depression between the inputs. To induce long-term synaptic modification, postsynaptic action potentials (APs) were paired with presynaptic stimulation of one input (60 pairs, 0.2 Hz). At the paired input, pre → post pairing (presynaptic stimulation followed by postsynaptic stimulation) induced long-term potentiation (LTP) and post → pre pairing (postsynaptic stimulation followed by presynaptic stimulation) induced long-term depression (LTD) (Fig. 1c), consistent with spike-timing-dependent plasticity (STDP) of these synapses¹⁴. Interestingly, there are quantitative differences in the temporal window for proximal and distal inputs (Fig. 1d). Although the magnitude of LTP was smaller distally than proximally, the main difference was in the width of the LTD window, with the distal window markedly broader than the proximal window. In particular, with a pre-/postsynaptic spike interval of $-100 < \Delta t < -50$ ms (Fig. 1d, dotted box), LTD was absent at proximal synapses ($-1.9 \pm 2.7\%$, $n = 8$; $P > 0.2$, t -test), but was still observed at distal synapses ($-22.6 \pm 3.4\%$, $n = 13$, $P < 0.00001$). This difference in the LTD window was also found with the GABA_A receptor antagonist picrotoxin (10 μM) in the bath (see Supplementary Fig. 1).

Induction of LTP by pre → post pairing may depend on supralinear summation of EPSPs and back-propagating APs^{9,15,18}. The

smaller magnitude of LTP at the distal dendrite may be due to attenuation of the AP^{9,18–21} over the length of the dendrite. For LTD induced by post → pre pairing, however, little is known about the underlying mechanism. We therefore performed a series of experiments examining the properties of distal and proximal dendrites that may account for the difference in the LTD window. We first measured the temporal interaction between EPSPs and back-propagating APs by making dendritic recordings either distally or proximally to measure the EPSP, the antidromically-evoked AP, and the paired AP and EPSP (Fig. 2a). In contrast to supralinear summation between EPSP → AP pairs^{9,15,18,21,22}, we found significant suppression of EPSPs in AP → EPSP pairs. This suppression had different temporal windows for distal and proximal inputs (Fig. 2b), reminiscent of the LTD windows (Fig. 1d). Such difference in the EPSP suppression window found in dendritic recordings (see Supplementary Fig. 2) was also detected in somatic recordings (Fig. 2c), a method used routinely in subsequent experiments.

The EPSP suppression in AP → EPSP pairs might be due to a reduction in presynaptic glutamate release (for example, caused by retrograde signals²³) or changes in postsynaptic responsiveness^{6,24–26}. When glutamate pulses were applied iontophoretically at each dendritic site, we found significant suppression of the responses

by the preceding AP (Fig. 2d), with temporal windows similar to those for EPSPs (Fig. 2c). This is consistent with a postsynaptic mechanism. We next examined whether the AP induces suppression of the synaptic response mediated by AMPARs (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors) or NMDA receptors (NMDARs). After isolation of AMPAR-mediated EPSPs (AMPA-EPSPs) using the selective NMDAR antagonist AP5 (D(-)-2-amino-5-phosphonopivalic acid, 50 μ M) and picrotoxin, we found that APs induced suppression of AMPAR-EPSPs, but with similar windows at distal and proximal dendrites (somatic recording, Fig. 2e; dendritic recording, Supplementary Fig. 3). This transient suppression may reflect shunting of AMPAR-EPSPs⁶. However, for NMDAR-EPSPs measured in the presence of CNQX (6-cyano-7-nitroquinoxaline-2,3-dione, 20 μ M) and picrotoxin, the suppression window was significantly broader distally than proximally (somatic recording, Fig. 2f; dendritic recording, Supplementary Fig. 3), resembling the windows for EPSP suppression with AP → EPSP pairing (Fig. 2b, c). These results suggest that suppression of NMDAR-EPSPs sets the spike-timing window for LTD.

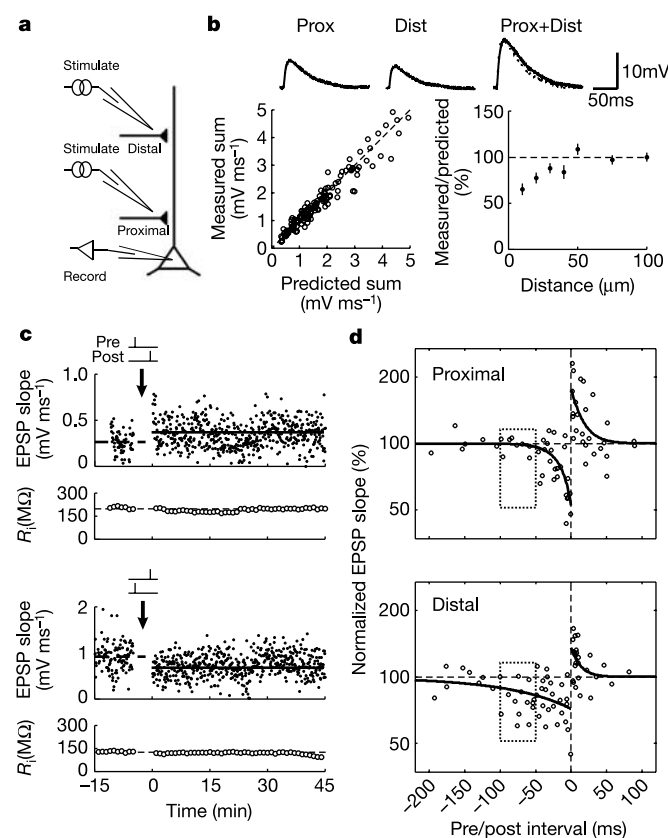


Figure 1 Spike-timing-dependent synaptic modification at proximal and distal dendrites. **a**, Experimental configuration. **b**, Focal stimulation activates separate distal and proximal inputs. Upper panel, recorded (solid) and predicted (dotted) EPSPs. Lower left, EPSP evoked by simultaneous stimulation versus linear prediction ($n = 110$). Lower right, measured/predicted EPSP ratio (mean $\pm$ s.e.m., $n = 4–13$) versus distance between stimulation sites. **c**, Top, example of LTP (EPSP slope: 48%; peak: 68%, not shown) of distal input induced by pre → post pairing ($\Delta t = 2$ ms). Lines indicate means before (dashed) and 11–20 min after (solid) induction (arrow). Bottom, example of LTD (slope: -22% ; peak: -19%) of distal input by post → pre pairing ($\Delta t = -52$ ms). **d**, Synaptic modification versus spike interval. Curves, single-exponential fits. Proximal: $A_+ = 0.76$, $\tau_+ = 15.9$ ms ($n = 26$); $A_- = -0.48$, $\tau_- = 19.3$ ms ($n = 38$). Distal: $A_+ = 0.36$, $\tau_+ = 12.5$ ms ($n = 24$); $A_- = -0.28$, $\tau_- = 103.4$ ms ($n = 46$).

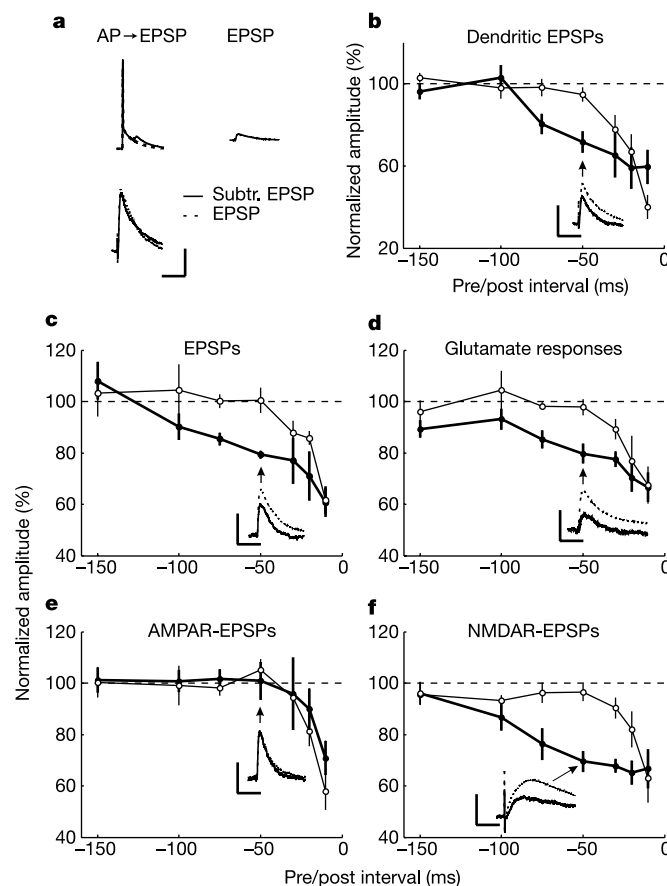


Figure 2 Suppression of EPSP by back-propagating action potential. **a**, Top, dendritic recordings (23 μ m from soma) of AP → EPSP pair (solid, $\Delta t = -46$ ms), AP (dashed), and control EPSP. Bottom, subtracted and control EPSPs; no suppression in this case. Scale: 40 mV (top), 4 mV (bottom); 100 ms. **b**, Temporal windows of AP → EPSP suppression in distal (filled circles) and proximal (open circles) dendrites. Subtracted EPSP amplitude normalized by control (mean $\pm$ s.e.m., $n = 4–9$) versus pre/post interval. Inset: subtracted ($\Delta t = -49$ ms) and control EPSPs recorded from distal dendrite (110 μ m). Scale: 2 mV; 100 ms. **c–f**, Similar to (**b**) except that recordings were made somatically. Distal inputs (filled circles); proximal inputs (open circles). AP-induced suppression of EPSPs (**c**), responses to glutamate pulses (**d**), AMPAR-EPSPs (**e**), and NMDAR-EPSPs (**f**) versus pre/post interval. $n = 4–12$. Scale: 4 mV (**c**) and 2 mV (**d–f**); 100 ms.

Dendritic recordings showed that suppression occurred in the absence of after-hyperpolarization (Fig. 2a and Supplementary Fig. 2a), indicating that it is unlikely to be caused by hyperpolarization-induced increase in Mg^{2+} block of NMDARs. Another possible mechanism is NMDAR desensitization^{24–26}, perhaps triggered by Ca^{2+} influx through AP-activated channels. To test this possibility, we measured AP-induced suppression of NMDAR-EPSPs after postsynaptic loading of BAPTA (5 mM), a high-affinity Ca^{2+} chelator, through the whole-cell pipette. As shown in Fig. 3a, BAPTA blocked the NMDAR-EPSP suppression, indicating that the suppression is Ca^{2+} -dependent. The suppression is also largely abolished by bath application of the L-type calcium channel antagonist nimodipine (10 μ M, Fig. 3b), indicating that Ca^{2+} influx through L-type channels is required. Finally, postsynaptic loading of calcineurin inhibitory peptide (CIP, 270 μ M, Fig. 3c) or infection of the cell with a viral vector expressing a peptide corresponding to the calcineurin-interacting domain of the NR2A subunit of NMDARs (see Supplementary Fig. 4a) largely abolished AP-induced NMDAR-EPSP suppression, indicating that calcineurin is required. Together, these results strongly suggest that AP-induced NMDAR-EPSP suppression is related to Ca^{2+} -dependent NMDAR desensitization^{24–26}.

To examine the relationship between NMDAR-EPSP suppression and LTD, we performed pharmacological experiments on LTD induction. We found that LTD induction by post $\rightarrow$ pre pairing depends on NMDARs^{13,17}, as it was blocked by AP5 (Fig. 3d). More

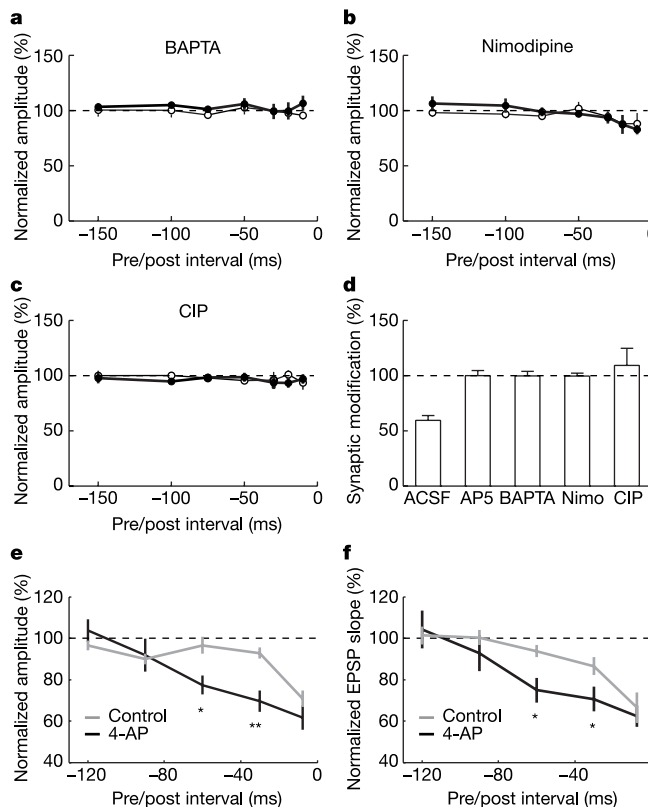


Figure 3 Pharmacological properties of NMDAR-EPSP suppression and LTD. **a–c**, AP-induced NMDAR-EPSP suppression for distal (filled circles) and proximal (open circles) inputs with BAPTA (**a**), nimodipine (**b**) or CIP (**c**); $n = 4–6$. **d**, LTD induced by post $\rightarrow$ pre pairing ($\Delta t = 6.4 \pm 3.3$ ms, s.d.) in normal ACSF and in presence of pharmacological agents ($n = 5–18$). Error bars show s.e.m. **e**, Temporal windows of AP-induced suppression of NMDAR-EPSPs at proximal dendrite with or without 4-aminopyridine (4-AP) ($n = 4–8$). Single asterisk, $P < 0.05$; double asterisk, $P < 0.01$; t -test. **f**, Temporal windows for LTD of proximal inputs induced by post $\rightarrow$ pre pairing with or without 4-aminopyridine ($n = 2–13$).

importantly, LTD was abolished by BAPTA, CIP, nimodipine^{13,17} (Fig. 3d), or viral expression of the NR2A peptide (see Supplementary Fig. 4b). We also measured dendritic Ca^{2+} elevation triggered by somatically elicited single APs using two-photon imaging, and found that Ca^{2+} signals were larger distally ($101 \pm 17 \mu$ m; $\pm$ s.d.) than proximally ($21 \pm 6 \mu$ m) (see Supplementary Fig. 5a, b), consistent with previous reports^{20,21}. These larger Ca^{2+} signals may cause longer suppression of NMDAR-EPSPs and thereby widen the LTD window. To examine further the relationship between AP-evoked dendritic Ca^{2+} signals and the windows for EPSP suppression and LTD induction, we bath-applied the K^+ channel blocker 4-aminopyridine (3 mM) to broaden the AP^{15,17} (see Supplementary Fig. 2c). We found that 4-aminopyridine caused a significant but reversible increase in AP-induced Ca^{2+} signals proximally (see Supplementary Fig. 5c). Importantly, 4-aminopyridine (in the recording pipette) also caused significant broadening of the proximal windows for NMDAR-EPSP suppression and LTD (Fig. 3e, f). Thus, the suppression and LTD windows co-vary with AP-induced dendritic Ca^{2+} signals, supporting the hypothesis that Ca^{2+} -dependent suppression of NMDAR-EPSPs

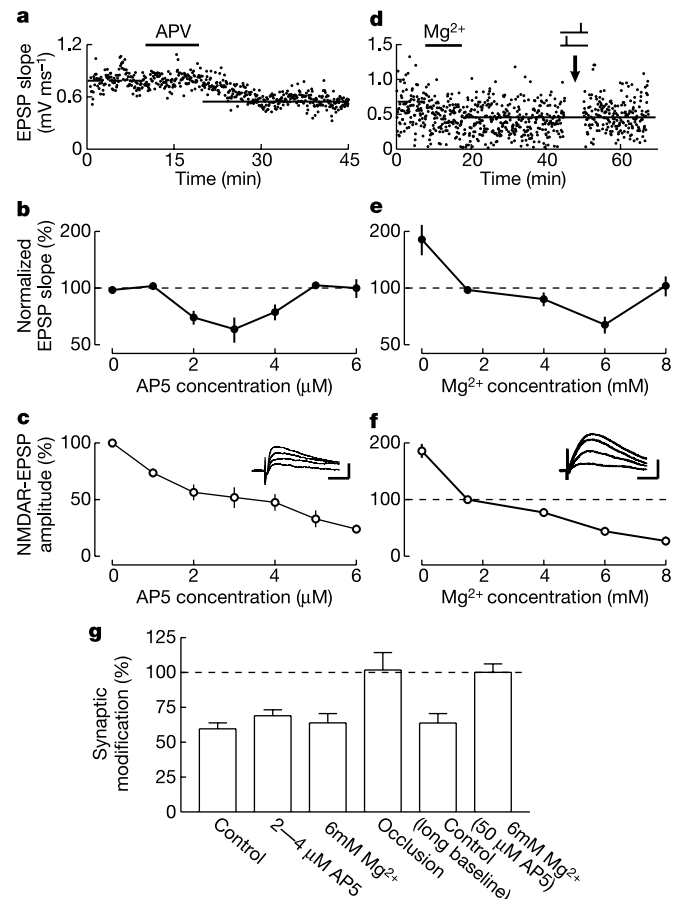


Figure 4 Synaptic modification induced by presynaptic stimulation under partial NMDAR blockade. **a**, Example LTD (-28.8%) induced by 2μ M AP5. Black bar shows AP5 wash-in (10 min for all experiments). **b**, LTD versus AP5 concentration. Error bars show s.e.m. ($n = 3–5$). **c**, NMDAR-EPSP versus AP5 concentration ($n = 4$). Inset, NMDAR-EPSPs (0, 1, 3, 5 μ M AP5). Scale: 0.5 mV; 35 ms. **d–f**, Similar to **a–c**, except with Mg^{2+} washed in. **d**, LTD following 6 mM Mg^{2+} wash-in: -32.4% . Post $\rightarrow$ pre pairing was applied 20 min after washout. **e**, $n = 3–5$. **f**, $n = 4–7$. Inset, NMDAR-EPSPs (1.5, 4, 6, 8 mM $[Mg^{2+}]_o$). Scale: 1 mV; 50 ms. **g**, Summary of LTD induced by AP5 and Mg^{2+} . Control: post $\rightarrow$ pre pairing; occlusion: post $\rightarrow$ pre pairing 20–30 min after washout of 6 mM $[Mg^{2+}]_o$; control (long baseline): post $\rightarrow$ pre pairing after 25–45 min of recording. Error bars show s.e.m. ($n = 3–18$).

determines the spike-timing window for LTD. A recent study in layer 5 of the visual cortex indicated that endocannabinoids serve as retrograde messengers for spike-timing-dependent LTD, and blocking their degradation widened the LTD window¹⁶. We note that, in addition to the potential existence of layer-specific mechanisms, the previous and current studies addressed different loci in the cellular pathways leading to LTD.

To determine whether reduction of NMDAR activation is sufficient for LTD, we partially blocked NMDARs using bath application of AP5 or by altering the bath concentration of Mg^{2+} ($[Mg^{2+}]_o$), instead of using postsynaptic APs. We found that presynaptic stimulation alone (0.2 Hz) induced long-lasting synaptic depression with low concentrations of AP5 (2–4 μM , $P < 0.05$), but not in the absence of or at high concentrations of AP5 (Fig. 4a–c, g). Similarly, presynaptic stimulation induced significant depression with 6 mM $[Mg^{2+}]_o$ ($P < 0.01$), but not with ≤ 4 or 8 mM $[Mg^{2+}]_o$ (Fig. 4d–g). Importantly, post $\rightarrow$ pre pairing induced LTD after 25–45 min of baseline recording, but post $\rightarrow$ pre pairing following 6 mM $[Mg^{2+}]_o$ -induced depression failed to induce LTD (Fig. 4d, g).

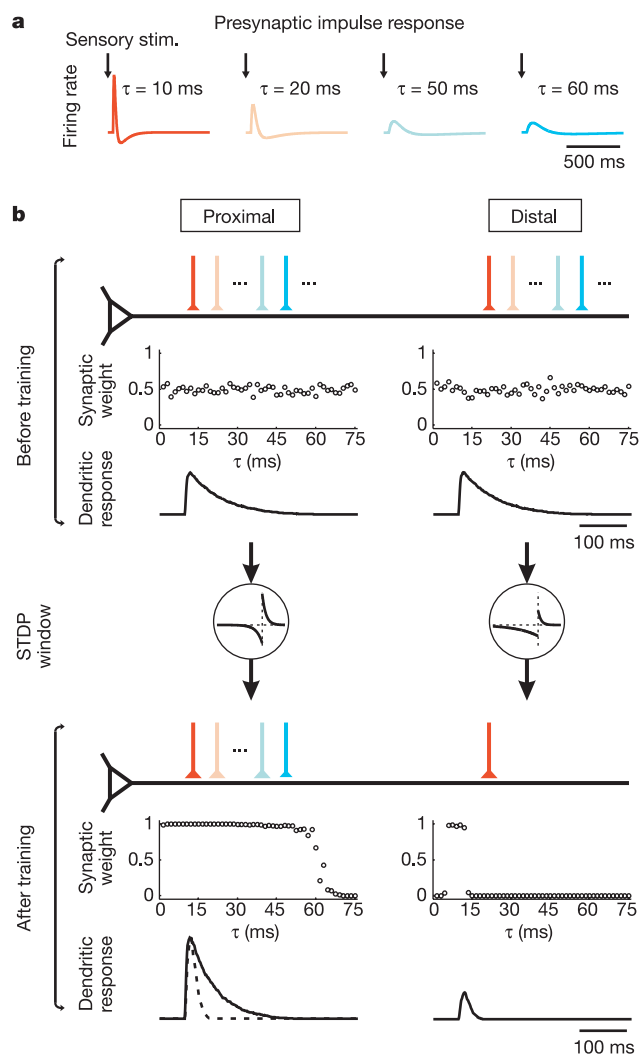


Figure 5 Simulation of location-dependent selection of synaptic inputs. **a**, Example presynaptic responses to brief sensory stimulus. Warmer colours, more transient responses. **b**, Synaptic inputs before and after 30 min of sensory stimuli. Shown are diagrams of input patterns, synaptic weights versus τ , and dendritic responses to brief sensory stimulus (see Supplementary Methods, equation (2)). To facilitate comparison of response time course after training, the distal dendritic response was scaled up (dashed line) to match the peak amplitude of the proximal response.

Therefore, Mg^{2+} -induced depression occluded pairing-induced LTD, suggesting common mechanisms for both processes. Furthermore, we found that Mg^{2+} -induced synaptic depression is NMDAR-dependent, as it is abolished by AP5 (Fig. 4g). Together, these results indicate that partial block of NMDARs is sufficient to induce LTD. Since the primary role of NMDAR activation in LTP/LTD induction is to control Ca^{2+} influx^{13,27}, our results support the hypothesis that AP-induced reduction of NMDAR-mediated Ca^{2+} entry (which may depend on cell type^{22,28}) leads to LTD, which fits well with the conventional model that links changes in intracellular Ca^{2+} with LTD induction^{13,17,27}. Interestingly, in 0 mM $[Mg^{2+}]_o$, which boosted NMDAR-EPSPs (Fig. 4f), presynaptic stimulation alone produced long-lasting EPSP enhancement (Fig. 4e, $P < 0.05$), consistent with the notion that enhanced Ca^{2+} influx through NMDARs induces LTP^{9,13,17,27}.

What are the functional consequences of the dendritic inhomogeneity in spike-timing-dependent synaptic modification? Previous theoretical studies show that LTD induced by post $\rightarrow$ pre pairing preferentially weakens inputs with long response latencies²⁹. The different LTD windows reported here (Fig. 1d) may lead to differential input selectivity along the apical dendrite. To test this idea, we constructed a simple model (see Methods) in which the proximal and distal dendrites of the postsynaptic neuron exhibit different STDP windows, based on our experimental data. This neuron receives multiple inputs driven by the same sensory stimuli, but with different temporal response characteristics (Fig. 5a). Starting with similar connection strengths for all inputs, we simulated presynaptic spike trains driven by random stimuli, then simulated the postsynaptic spike train by integrating all synaptic inputs and modified each connection according to the STDP windows. We found that transient inputs were strengthened and sustained inputs were weakened both distally and proximally, but the distal dendrite was much more selective (Fig. 5b). Comparison of the dendritic responses to a brief sensory stimulus showed that, although initially similar, the response became much more transient at distal than at proximal dendrites after training. Thus, the different STDP windows may lead to functional differentiation of the distal and proximal dendrites in receiving signals with distinct temporal characteristics. Owing to the selection for transient inputs, distal dendrites may be specialized for processing the precise timing of sensory signals.

Spike-timing-dependent plasticity is a robust learning rule in the central nervous system^{9–17}. Although the basic asymmetry in the spike-timing window is found at many synapses, the width of the window, especially that for LTD induction, can be highly variable¹¹. Our findings indicate that such variability may be attributed in part to differences in the dendritic location of the synapses and in the temporal dynamics of AP-induced Ca^{2+} elevation that leads to NMDAR desensitization. More importantly, the location-specific STDP windows provide a mechanism for developmental segregation of inputs carrying different sensory signals. Although spatial inhomogeneity in the electrical properties and nonlinearity in dendritic integration endow individual pyramidal neurons with enhanced processing capacity^{5,8}, the computational power can only be harvested if inputs carrying different signals are parsed into distinct dendritic regions^{8,30}. The present findings not only demonstrate dendritic inhomogeneity in activity-dependent synaptic modification, but also point to its potential role in the formation of location-specific synaptic inputs, a feature that could greatly enrich the repertoire of dendritic information processing. □

Methods

Visual cortical slice preparation

Acute visual cortical slices were prepared from 2–4-week-old Sprague–Dawley rats. Rats were deeply anaesthetized with halothane, decapitated, and the brain quickly placed into ice-cold dissection buffer containing 206 mM sucrose, 2.5 mM KCl, 2 mM $MgSO_4$,

1.25 mM NaH₂PO₄, 1 mM CaCl₂, 1 mM MgCl₂, 26 mM NaHCO₃ and 10 mM dextrose, bubbled with 95% O₂/5% CO₂ (pH 7.4). Slices (300–400 µm thick) were prepared with a vibratome (Pelco), placed in warm dissection buffer (33–35 °C) for <30 min, transferred to a holding chamber containing artificial cerebrospinal fluid (ACSF: 124 mM NaCl, 2 mM KCl, 1.5 mM MgSO₄, 1.25 mM NaH₂PO₄, 2.5 mM CaCl₂, 26 mM NaHCO₃, 10 mM dextrose), and kept at 22–24 °C for >1 h before use. For experiments, slices were transferred to the recording chamber and perfused (4.0–4.5 ml min⁻¹) with oxygenated ACSF at 22–24 °C.

Electrophysiology

Somatic and dendritic whole-cell recordings were made in current-clamp with an Axopatch 200B amplifier (Axon) using infrared differential interference optics (IR-DIC) video microscopy. Layer 2/3 pyramidal cells were selected based on morphological and electrophysiological criteria^{12,14}. Patch pipettes (somatic: 3–8 MΩ; dendritic: 8–20 MΩ) were filled with intracellular solution (120 mM K-gluconate, 10 mM HEPES, 0.1 mM EGTA, 20 mM KCl, 2 mM MgCl₂, 10 mM phosphocreatine, 2 mM ATP, and 0.25 mM GTP). The mean resting potential was -70.3 ± 0.7 mV, corrected for the measured liquid junction potential (6.8 mV). The series resistance was 14.6 ± 1.6 MΩ. Input resistance ($R_i = 118.0 \pm 5.4$ MΩ) was monitored with hyperpolarizing current pulses (50 pA, 100 ms); cells were excluded if R_i changed >30% over the entire experiment^{12,14}. Data were filtered at 2 kHz, digitized at 10 kHz and analysed with Clampfit 8 (Axon). EPSPs were evoked by focal extracellular stimulation (0.01–1 ms, 1–100 V) with small glass bipolar electrodes. To ensure that the distal (>100 µm from soma) and proximal (<50 µm) electrodes activated separate synapses, we tested both the linear summation of EPSPs (Fig. 1b) and paired-pulse depression between inputs. Although consecutive stimulation through the same electrode induced marked paired-pulse depression of these synapses, no depression was induced by sequential stimulation through the two electrodes, indicating that they activated separate synapses. Postsynaptic APs were elicited either with depolarizing current injection through the recording electrode (1 nA, 1–4 ms; somatic recordings) or antidromically with an extracellular stimulating electrode near the axonal initial segment (0.01–1 ms, 5–100 V; dendritic recordings). Presynaptic spike timing was defined as the onset of EPSP and postsynaptic spike timing was measured at the AP peak^{12,14}. Synaptic strength was measured as the initial slope (first 2 ms) of the EPSP. To measure long-term synaptic modification, a stable baseline of synaptic strength was first established by 6–12 min of recording with presynaptic stimulation at 0.2 Hz. Synaptic strength after induction was measured 11–20 min after the end of induction. For iontophoretic application of glutamate, a sharp microelectrode (150–200 MΩ) filled with 250 mM Na⁺-glutamate was positioned near the apical dendrite (<5 µm). Both the holding current (1–10 nA) and ejection current (100–300 nA, 0.01–3 ms) were applied with an amplifier (Getting)³.

Model simulation

The model consisted of one postsynaptic neuron and 100 presynaptic inputs (50 to the proximal and 50 to the distal dendrite, see diagram in Fig. 5b). The presynaptic neurons had a range of response time courses (Fig. 5a) but they were all driven by the same sensory stimulus, which was a temporally varying random signal (see Supplementary Fig. 6a). The spike train of the postsynaptic cell was simulated after integrating the synaptic input to both the distal and proximal dendritic compartments. All connection weights were initialized to 0.5 plus a small random number (Fig. 5b, upper panel), and were modified according to the STDP windows measured experimentally. Details of the model are given in Supplementary Methods.

Received 11 May; accepted 23 December 2004; doi:10.1038/nature03366.

- Stuart, G. & Spruston, N. Determinants of voltage attenuation in neocortical pyramidal neuron dendrites. *J. Neurosci.* **18**, 3501–3510 (1998).
- Larkum, M. E., Zhu, J. J. & Sakmann, B. A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature* **398**, 338–341 (1999).
- Cash, S. & Yuste, R. Linear summation of excitatory inputs by CA1 pyramidal neurons. *Neuron* **22**, 383–394 (1999).
- Magee, J. C. Dendritic integration of excitatory synaptic input. *Nature Rev. Neurosci.* **1**, 181–190 (2000).
- Segev, I. & London, M. Untangling dendrites with quantitative models. *Science* **290**, 744–750 (2000).
- Hausser, M., Major, G. & Stuart, G. J. Differential shunting of EPSPs by action potentials. *Science* **291**, 138–141 (2001).
- Tamas, G., Szabadics, J. & Somogyi, P. Cell type- and subcellular position-dependent summation of unitary postsynaptic potentials in neocortical neurons. *J. Neurosci.* **22**, 740–747 (2002).
- Hausser, M. & Mel, B. Dendrites: bug or feature? *Curr. Opin. Neurobiol.* **13**, 372–383 (2003).
- Magee, J. C. & Johnston, D. A. synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. *Science* **275**, 209–213 (1997).
- Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215 (1997).
- Sourdet, V. & Debanne, D. The role of dendritic filtering in associative long-term synaptic plasticity. *Learn. Mem.* **6**, 422–447 (1999).
- Feldman, D. E. Timing-based LTP and LTD at vertical inputs to layer II/III pyramidal cells in rat barrel cortex. *Neuron* **27**, 45–56 (2000).
- Bi, G. & Poo, M. Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.* **24**, 139–166 (2001).
- Frommke, R. C. & Dan, Y. Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature* **416**, 433–438 (2002).
- Watanabe, S., Hoffman, D. A., Migliore, M. & Johnston, D. Dendritic K⁺ channels contribute to spike-timing dependent long-term potentiation in hippocampal pyramidal neurons. *Proc. Natl Acad. Sci. USA* **99**, 8366–8371 (2002).

- Sjostrom, P. J., Turrigiano, G. G. & Nelson, S. B. Neocortical LTD via coincident activation of presynaptic NMDA and cannabinoid receptors. *Neuron* **39**, 641–654 (2003).
- Johnston, D. et al. Active dendrites, potassium channels and synaptic plasticity. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **358**, 667–674 (2003).
- Stuart, G. J. & Hausser, M. Dendritic coincidence detection of EPSPs and action potentials. *Nature Neurosci.* **4**, 63–71 (2001).
- Golding, N. L., Staff, N. P. & Spruston, N. Dendritic spikes as a mechanism for cooperative long-term potentiation. *Nature* **418**, 326–331 (2002).
- Svoboda, K., Helmchen, F., Denk, W. & Tank, D. W. Spread of dendritic excitation in layer 2/3 pyramidal neurons in rat barrel cortex *in vivo*. *Nature Neurosci.* **2**, 65–73 (1999).
- Waters, J., Larkum, M., Sakmann, B. & Helmchen, F. Supralinear Ca²⁺ influx into dendritic tufts of layer 2/3 neocortical pyramidal neurons *in vitro* and *in vivo*. *J. Neurosci.* **22**, 8558–8567 (2003).
- Koester, H. J. & Sakmann, B. Calcium dynamics in single spines during coincident pre- and postsynaptic activity depend on relative timing of back-propagating action potentials and subthreshold excitatory postsynaptic potentials. *Proc. Natl Acad. Sci. USA* **95**, 9596–9601 (1998).
- Zilberter, Y., Kaiser, K. M. & Sakmann, B. Dendritic GABA release depresses excitatory transmission between layer 2/3 pyramidal and bitufted neurons in rat neocortex. *Neuron* **24**, 979–988 (1999).
- Rosenmund, C., Feltz, A. & Westbrook, G. L. Calcium-dependent inactivation of synaptic NMDA receptors in hippocampal neurons. *J. Neurophysiol.* **73**, 427–430 (1995).
- Tong, G., Shepherd, D. & Jahr, C. E. Synaptic desensitization of NMDA receptors by calcineurin. *Science* **267**, 1510–1512 (1995).
- Umeyama, M., Chen, N., Raymond, L. A. & Murphy, T. H. A calcium-dependent feedback mechanism participates in shaping single NMDA miniature EPSCs. *J. Neurosci.* **21**, 1–9 (2001).
- Zucker, R. S. Calcium- and activity-dependent synaptic plasticity. *Curr. Opin. Neurobiol.* **9**, 305–313 (1999).
- Nevian, T. & Sakmann, B. Single spine Ca²⁺ signals evoked by coincident EPSPs and backpropagating action potentials in spiny stellate cells of layer 4 in the juvenile rat somatosensory barrel cortex. *J. Neurosci.* **24**, 1689–1699 (2004).
- Song, S., Miller, K. D. & Abbott, L. F. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neurosci.* **3**, 919–926 (2000).
- Archie, K. A. & Mel, B. W. A model for intradendritic computation of binocular disparity. *Nature Neurosci.* **3**, 54–63 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Ascher, G. Bi, L. Chen, M. Frerking, E. Isacoff, R. Kramer and R. Zucker for helpful discussions, and K. Arendt, K. Borges, N. Caporale and C. Nam for technical assistance. This work was supported by grants from the National Eye Institute and the Grass Foundation. R.C.F. is a recipient of the Howard Hughes Predoctoral Fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.D. (ydan@berkeley.edu).

The receptors and coding logic for bitter taste

Ken L. Mueller¹, Mark A. Hoon², Isolde Erlenbach², Jayaram Chandrashekar¹, Charles S. Zuker¹ & Nicholas J. P. Ryba²

¹Howard Hughes Medical Institute and Departments of Biology and Neurosciences, University of California at San Diego, La Jolla, California 92093-0649, USA

²National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, Maryland 20892, USA

The sense of taste provides animals with valuable information about the nature and quality of food. Bitter taste detection functions as an important sensory input to warn against the ingestion of toxic and noxious substances. T2Rs are a family of approximately 30 highly divergent G-protein-coupled receptors (GPCRs)^{1,2} that are selectively expressed in the tongue and palate epithelium¹ and are implicated in bitter taste sensing^{1–8}. Here we demonstrate, using a combination of genetic, behavioural and physiological studies, that T2R receptors are necessary and sufficient for the detection and perception of bitter compounds, and show that differences in T2Rs between species (human and mouse) can determine the selectivity of bitter taste responses. In addition, we show that mice engineered to express a bitter taste

receptor in 'sweet cells' become strongly attracted to its cognate bitter tastants, whereas expression of the same receptor (or even a novel GPCR) in T2R-expressing cells resulted in mice that are averse to the respective compounds. Together these results illustrate the fundamental principle of bitter taste coding at the periphery: dedicated cells act as broadly tuned bitter sensors that are wired to mediate behavioural aversion.

Two main lines of evidence suggest that T2Rs function as mammalian bitter taste receptors. First, heterologous expression assays have shown that several human and rodent T2Rs respond to bitter tasting compounds: mouse (m)T2R5 is a high affinity receptor for cycloheximide (Cyx)³, human (h)T2R16 is a candidate receptor for β -glucopyranosides (salicin and related compounds)⁴, hT2R14 is a candidate receptor for picrotoxinin⁶, and hT2R44 and hT2R61/hT2R43 are receptors for denatonium, aristolochic acid and 6-nitrosaccharin^{7,8}. Second, sequence polymorphisms in T2Rs have been linked to differences in bitter taste sensitivity in mice³ and humans⁵. To determine if T2Rs and T2R-expressing cells are necessary and sufficient for bitter taste perception *in vivo*, we used several complementary strategies. First, we genetically engineered mice to express candidate human T2R receptors for tastants that mice do not respond to, and examined whether introduction of these candidate bitter receptors endows the animals with an expanded bitter taste repertoire. Second, we generated genetically modified mice lacking a specific T2R and studied their behavioural

and physiological responses to bitter compounds. Third, we examined the specificity of bitter taste responses of animals in which all sweet, umami and bitter taste function was eliminated and then selectively restored in T2R-expressing cells. Finally, we functionally dissected the role of cells and receptors by ectopically expressing a T2R bitter taste receptor in sweet-sensing⁹ (T1R) cells.

Mice and humans have distinctive differences in their sensitivities to many bitter compounds. For example, several β -glucopyranosides evoke strong bitter taste in humans, yet mice are largely indifferent to these compounds. Similarly, phenylthiocarbamide (PTC), a well known bitter tastant often used in human genetic studies, is ineffective in mice (Fig. 1). We placed the candidate human receptor for β -glucopyranosides (hT2R16, ref. 4) and PTC (hT2R38, ref. 5) under the control of a mouse T2R promoter^{1,10}, and generated transgenic animals expressing either of these two candidate taste receptors in T2R-expressing cells. To examine taste behaviour, we measured taste choices in two-bottle intake preference assays, or by direct counting of immediate licking responses in a multi-channel gustometer (see Methods and ref. 10). Figure 1 shows that T2R-hT2R16 transgenic animals acquire the ability to detect and respond to phenyl- β -D-glucopyranoside at concentrations that closely approximate the human physiological taste sensitivity range⁴. Likewise, expression of hT2R38 confers selective PTC sensitivity to the engineered mice. These results validate T2Rs and T2R-expressing cells as mediators of bitter taste perception *in vivo*, and demonstrate that T2Rs are sufficient for selective responses to bitter tastants; they also confirm that hT2R16 is the β -glucopyranoside receptor⁴ and hT2R38 is the PTC receptor⁵. In addition, the ability to confer human bitter taste responses on mice by introduction of human taste receptors illustrates an important feature of T2Rs and bitter taste: selectivity and sensitivity differences to bitter compounds between species is probably a reflection of sequence differences in the respective T2R repertoires.

Next, we used homologous recombination to generate animals lacking mT2R5, the candidate Cyx receptor³. If T2R5 is the principal bitter taste receptor for Cyx, then its knockout should abolish most, if not all, responses to cycloheximide. To explore the effect of this knockout *in vivo*, we recorded tastant-induced action potentials from one of the principal nerves innervating taste receptor cells of the tongue. Figure 2 shows that T2R5^{-/-} mice have a dramatic and selective loss of responses to Cyx. In addition, the animals are no longer behaviourally averse to Cyx, even at concentrations 100-fold higher than those required to trigger strong repulsion in wild-type mice. As expected, responses to sweet, umami, sour and salty tastants are physiologically and behaviourally comparable to controls. To further examine the taste repertoire of the T2R5^{-/-} mice, we performed studies of bitter taste against a broad panel of bitter tastants. T2R5^{-/-} animals retain basically normal responses to all other bitter tastants tested (Fig. 2 and Supplementary Fig. 2). These results prove the essential requirement of a T2R receptor for bitter taste, and together with our T2R mis-expression studies (Fig. 1) show that defined T2Rs are both necessary and sufficient for bitter taste sensation. Interestingly, T2R5^{-/-} mice show residual responses to millimolar levels of cycloheximide, but mice defective in all bitter taste signalling (for example, phospholipase C β 2 knockout (PLC β 2^{-/-}) mice, ref. 10) have a complete loss of Cyx sensitivity (Fig. 2b); we suggest that this residual activity in T2R5^{-/-} mice reflects the recruitment of lower affinity T2Rs.

Previously, we showed that most T2R receptor genes are co-expressed in the same subset of taste receptor cells of the tongue and palate epithelium¹. We interpreted this to mean that individual T2R-expressing cells act as broadly tuned bitter sensors capable of responding to a wide diversity of tastants but not necessarily able to discriminate between them. Recently, we used mice deficient in sweet, umami and bitter taste (owing to a deletion of the PLC β 2 effector molecule¹⁰) to ask whether taste receptor cells are tuned to single or multiple taste modalities. By selectively rescuing PLC β 2

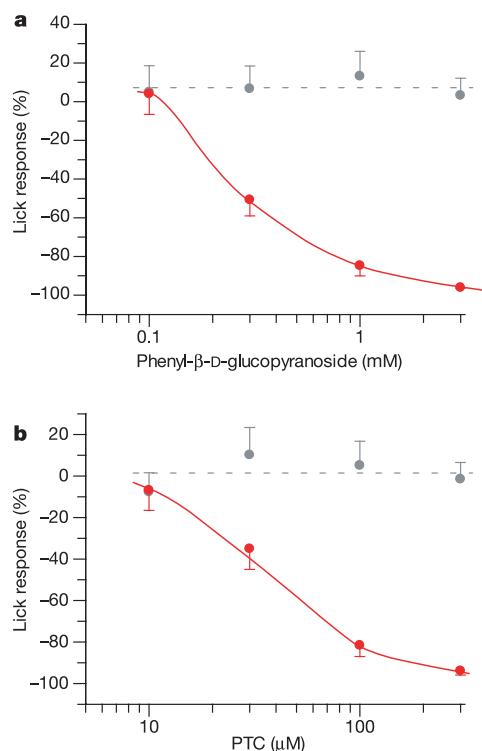


Figure 1 Introduction of human bitter receptors expands the bitter taste repertoire of mice. Brief-access taste tests measuring immediate lick responses show that normal FVB/N mice, unlike humans^{4,5}, do not taste phenyl- β -D-glucopyranoside or PTC (grey). In contrast, transgenic mice expressing the human T2R taste receptors for phenyl- β -D-glucopyranoside or PTC in bitter-sensing cells show robust behavioural aversion to these tastants (red). **a**, Mice expressing the human T2R16 taste receptor under the control of the mT2R19 promoter (red) respond to concentrations of phenyl- β -D-glucopyranoside that taste bitter to humans. **b**, Transgenic mice expressing human T2R38 (red) are averse to PTC at concentrations closely mimicking human sensitivity to this tastant. These mice respond normally to control bitter compounds (data not shown). Values are means $\pm$ s.e.m. ($n = 8$).

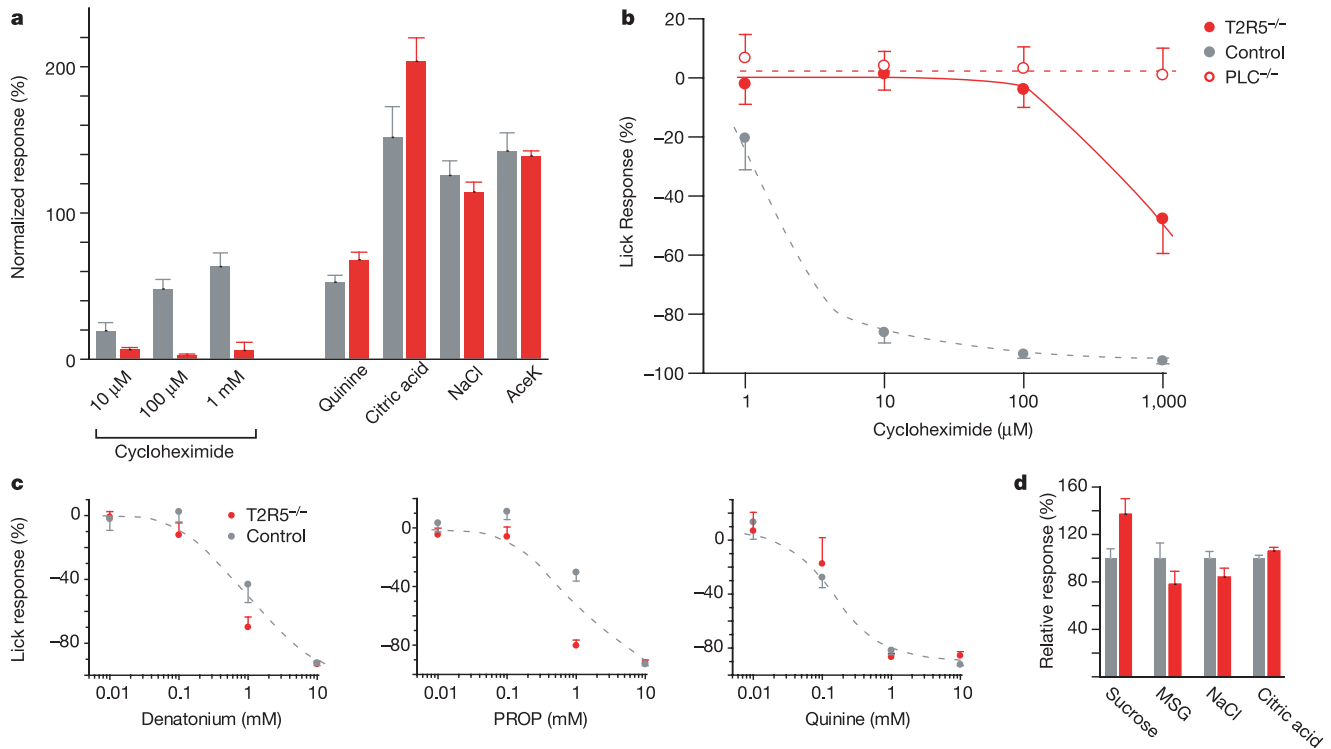


Figure 2 T2R5 is necessary for cycloheximide taste reception and perception. T2R5^{-/-} mice show a strong and selective impairment in their ability to taste cycloheximide. **a**, Electrophysiological responses to cycloheximide are essentially eliminated in knockout mice (red bars), but responses to other tastants are equivalent to those of control animals¹⁰ (grey bars; see Supplementary Information). Shown are integrated chorda tympani responses normalized to the response to 100 mM NH₄Cl (ref. 21). T2R5^{-/-} mice show a severe selective deficit in their ability to taste cycloheximide (**b**) but not to other

bitter (**c**) or sweet, umami, salty or sour tastants (**d**). **b**, At concentrations 100-fold higher than those needed to trigger maximal aversion in control animals, T2R5^{-/-} mice begin to show behavioural aversion. This probably reflects the recruitment of a low-sensitivity receptor for cycloheximide; total ablation of the bitter taste system (for example, PLC β 2^{-/-} animals; open circles) eliminates this residual response. Values are means $\pm$ s.e.m.; $n = 3$ (**a**), $n = 8$ (**b–d**). See Supplementary Information for responses to additional tastants in T2R5^{-/-} animals.

function only in T2R-expressing cells, we showed that there is complete functional segregation between attractive (sweet, umami) and aversive (bitter) tastes¹⁰. We now make use of a similar strategy to ask whether re-introduction of PLC β 2 under the control of selective T2R-promoters restores complete or only partial bitter

taste. If indeed individual T2R-cells express most bitter taste receptors, then targeting expression of a PLC β 2 'rescue' construct under the control of any one T2R promoter should be sufficient to restore normal bitter taste (that is, if the cell can now signal, all the receptors will function). Conversely, if different T2R-expressing

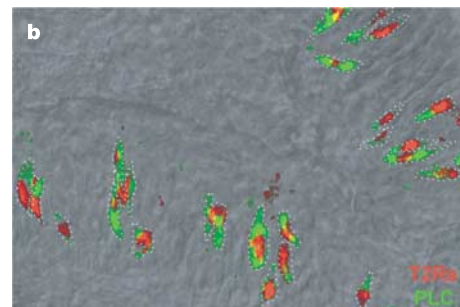
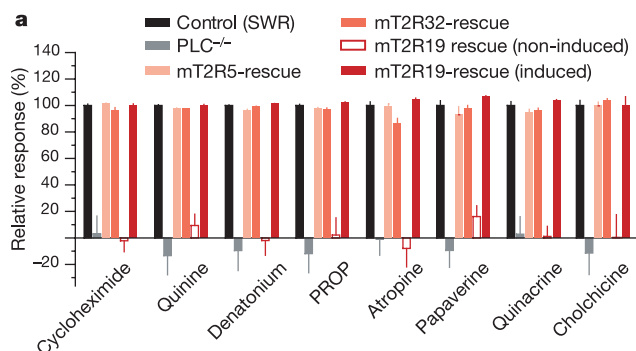


Figure 3 T2R-expressing cells are the mediators of bitter taste. **a**, Taste preferences of control (black), PLC β 2^{-/-} (grey), and transgenic 'rescue' lines that express PLC β 2 under the control of mT2R5, mT2R32 or mT2R19 (shades of red) were measured relative to water using a brief-access test against a panel of bitter tastants. The rescue line mT2R19 was engineered as a two component Tet-on system⁹. In the absence of doxycycline, the mT2R19 rescue mice (open red bars) do not taste bitter compounds. PLC β 2 knockout completely eliminates bitter taste behaviour¹⁰. However, expression of PLC β 2 under the control of each of the three different T2R promoters restores normal bitter taste to a broad panel of chemically diverse bitter tastants; this

demonstrates broadly tuned bitter sensing for the cells that express T2R5, T2R19 and T2R32, and demonstrates that a large repertoire of T2Rs is expressed in each T2R-expressing cell¹. Values shown are means $\pm$ s.e.m. ($n = 6–12$). For additional data and tastants, see Supplementary Fig. 3 **b**, T2R regulatory sequences drive PLC β 2 expression in T2R-expressing cells. Shown is double labelling of T2R32 driving PLC β 2 expression in a PLC β 2^{-/-} background, demonstrating coexpression of PLC β 2 (immunoreactivity, green) and T2Rs (*in situ* hybridization, red); dotted lines highlight taste receptor cells. The *in situ* signal is concentrated around the cell nucleus, whereas immunostaining is primarily cytoplasmic.

cells are narrowly tuned to different subsets of bitter tastants¹¹, then mice with restricted PLC expression would regain bitter sensitivity to a discrete repertoire of bitter-tasting compounds.

To address this question we chose three divergent T2Rs mapping to different chromosomal locations (T2R5, T2R19 and T2R32, ref. 1) and used their regulatory sequences to drive PLC β 2 expression in PLC β 2^{-/-} lines. We then tested the engineered animals against a broad panel of chemically diverse bitter compounds (Fig. 3 and Supplementary Fig. 3). In addition, we designed one of the three constructs to have an inducible expression system, in order to determine whether T2R signalling plays a critical role in the development and/or connectivity of T2R cells. Remarkably, all of the PLC β 2-rescued animals showed completely normal bitter taste (Fig. 3), including those in which T2R function was restored only at the adult stage, long after the taste system would have completed its normal development and wiring programme¹². These results show that the bitter taste circuitry can be established without bitter sensory input, and unequivocally demonstrate that individual T2R cells operate as broadly tuned bitter sensors.

Recently, we generated mice expressing a RASSL¹³ κ -opioid receptor in sweet cells and showed that these animals become selectively attracted to the synthetic opioid agonist spiradoline, a normally tasteless compound⁹. Thus, activation of sweet cells, rather than the sweet taste receptor itself, results in the perception of sweetness. Does the same logic apply to bitter taste? We tested this

idea by generating mice in which an inducible RASSL receptor was now targeted to bitter taste cells. Figure 4 shows that non-induced animals, or wild-type controls treated with doxycycline, are completely insensitive to the RASSL agonist spiradoline, even at millimolar concentrations. In contrast, mice expressing RASSL in bitter cells show strong aversion to spiradoline. Together, these results substantiate the coding of both sweet and bitter pathways by dedicated (that is, labelled) lines. A final corollary emerging from these findings is that expression of a sweet receptor in bitter cells should trigger behavioural aversion to sweet tastants, and expression of a bitter receptor in sweet cells should result in attraction to the bitter compound. Accordingly, we engineered mice expressing the bitter receptor for β -glucopyranosides in sweet cells. Indeed, these mice now display strong attraction to this family of bitter compounds. Thus, the 'taste' of a sweet or a bitter compound (that is, the perception of sweet and bitter) is a reflection of the selective activation of T1R-expressing versus T2R-expressing cells, rather than a property of the receptors or even the tastant molecules. □

Methods

Gene targeting of T2R5

T2R5 knockout mice were generated by homologous recombination following standard procedures^{9,10}. The entire coding sequence of T2R5 was replaced by a reverse-tetracycline dependent transactivator (rtTA) and a loxP-flanked PGK-neo^r cassette (see Supplementary Fig. 1 for details). Homologous recombination in R1 embryonic stem (ES) cells was detected by diagnostic Southern hybridization with probes outside the targeting construct. Two targeted ES clones were injected into C57BL/6 blastocysts. Chimaeric mice were bred with C57BL/6 mice and progeny backcrossed to C57BL/6 mice for two generations before establishing a homozygous knockout colony.

Transgenic animals

Transgenic lines were produced by pronuclear injection of zygotes from FVB/N or CB6 (BALB/c $\times$ C57BL/6 hybrids) mice. For constructs using the Tet-on inducible system, tetracycline-dependent gene expression was induced by feeding animals a diet containing doxycycline (6 g kg⁻¹) (Bio-Serv) for 3 days before and during behavioural testing⁹. For each construct, at least two independent animal lines were generated. The T2R38 transgenic construct used the PAV-taster allele⁵. For transgenic constructs using T2R regulatory sequences, we used the following fragments: T2R5, -10816 to +3 (the location of the ATG start codon)¹⁰; T2R19, -11012 to +3; T2R32, -9506 to +3. *In situ* hybridization and immunohistochemistry were carried out as described previously¹⁰.

Behavioural assays

Taste behaviour was assayed using a short-term assay that directly measures taste preferences by counting immediate licking responses in a multi-channel gustometer (Davis MS160-Mouse gustometer; DiLog Instruments)¹⁴. Before training and behavioural testing, all mice with a T2R5^{-/-} or PLC β 2^{-/-} background were treated with intranasal zinc sulphate¹⁵ to reduce input through the olfactory system. Mice were trained and tested as described previously^{9,10}. Salt attraction to 150 mM NaCl was measured in mice that had been salt deprived overnight¹⁶. Lick response represents the mean percentage rate at which mice licked a tested compound relative to their sampling of an appropriate control tastant; relative responses were scaled to the mean response of control animals. The concentrations of tastants used for bar graphs were: 300 mM sucrose; 100 mM glutamate + 1 mM inosine monophosphate + 0.1 mM amiloride (MSG); 150 mM NaCl (attraction); 150 mM citric acid; 10 μ M cycloheximide; 10 mM quinine; 10 mM denatonium; 10 mM 6-*n*-propyl-2-thiouracil (PROP); 10 mM papaverine; 10 mM quinacrine; 1 mM cholic acid, 5 mM atropine.

Standard two-bottle preference assays were carried out as described previously¹⁷. For the two-bottle assays in Fig. 4b, consumption relative to total is defined as intake of tastant divided by total intake (tastant plus water). For mice carrying rtTA and TetO-transgenes, controls included testing the same mice without induction as well as mice carrying just the rtTA transgene and exposed to doxycycline.

Nerve recordings

Lingual stimulation and recording procedures were performed as previously described^{18,19}. Neural signals were amplified (5,000 $\times$) with a Grass P511 AC amplifier (Astro-Med), digitized with a Digidata 1200B A/D converter (Axon Instruments), and integrated with a time constant of 0.5 s. Taste stimuli were presented at a constant flow rate of 4 ml min⁻¹ for 20 s intervals, interspersed by 2 min rinses with artificial saliva²⁰ between presentations. All data analyses used the integrated response over a 25 s period immediately after the application of the stimulus. The mean response to 100 mM NH₄Cl was used to normalize responses to each experimental series.

Tastants used for nerve recordings (maximal concentrations) were: 60 mM acesulfameK (AceK); 100 mM citric acid; 100 mM NaCl; 100 mM NH₄Cl; 10 mM quinine; 1 mM cycloheximide.

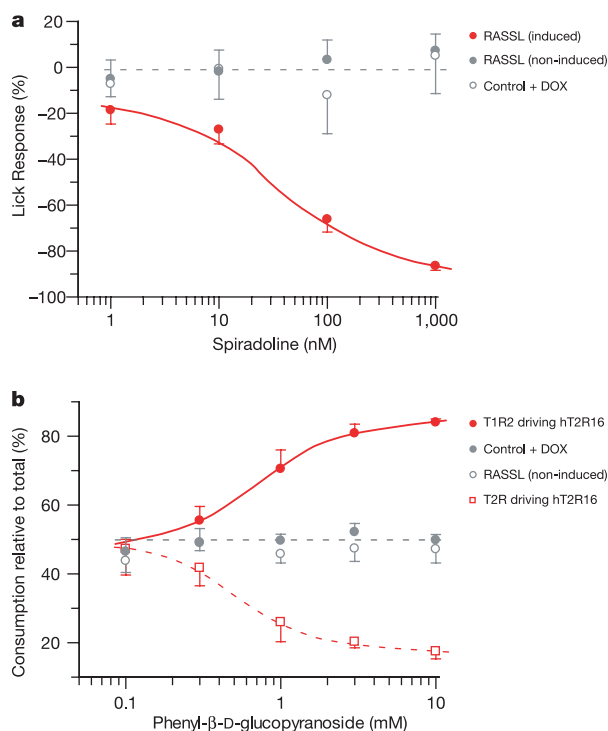


Figure 4 Switching the behavioural taste responses of mice by mis-expression of a bitter taste receptor. **a**, Expression of RASSL¹³ in T2R-expressing cells generates animals that show specific behavioural aversion to spiradoline (red). No responses are seen in control mice that carry the rtTA transgene but lack the RASSL receptor (open grey circles) or in non-induced RASSL mice (filled grey circles), even at 100 times the concentration needed to elicit strong responses in RASSL-expressing animals. **b**, Transgenic mice expressing human T2R16 (under control of the mT2R19 promoter) in bitter cells show strong behavioural aversion to phenyl- β -D-glucopyranoside (open red squares). Expression of the same receptor in sweet cells (T1R2-expressing cells) generates animals that show strong attraction for this bitter tastant (filled red circles). Control animals (filled grey circles) or non-induced experimental animals (open grey circles) show no preference or aversion to this tastant. Values are means $\pm$ s.e.m. ($n = 8$).

Received 3 December 2004; accepted 10 January 2005; doi:10.1038/nature03352.

- Adler, E. *et al.* A novel family of mammalian taste receptors. *Cell* **100**, 693–702 (2000).
- Matsunami, H., Montmayeur, J. P. & Buck, L. B. A family of candidate taste receptors in human and mouse. *Nature* **404**, 601–604 (2000).
- Chandrasekhar, J. *et al.* T2Rs function as bitter taste receptors. *Cell* **100**, 703–711 (2000).
- Bufe, B., Hofmann, T., Krautwurst, D., Raguse, J. D. & Meyerhof, W. The human TAS2R16 receptor mediates bitter taste in response to beta-glucopyranosides. *Nature Genet.* **32**, 397–401 (2002).
- Kim, U. K. *et al.* Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. *Science* **299**, 1221–1225 (2003).
- Behrens, M. *et al.* The human taste receptor hTAS2R14 responds to a variety of different bitter compounds. *Biochem. Biophys. Res. Commun.* **319**, 479–485 (2004).
- Pronin, A. N., Tang, H., Connor, J. & Keung, W. Identification of ligands for two human bitter T2R receptors. *Chem. Senses* **29**, 583–593 (2004).
- Kuhn, C. *et al.* Bitter taste receptors for saccharin and acesulfame K. *J. Neurosci.* **24**, 10260–10265 (2004).
- Zhao, G. Q. *et al.* The receptors for mammalian sweet and umami taste. *Cell* **115**, 255–266 (2003).
- Zhang, Y. *et al.* Coding of sweet, bitter, and umami tastes: different receptor cells sharing similar signaling pathways. *Cell* **112**, 293–301 (2003).
- Caicedo, A. & Roper, S. D. Taste receptor cells that discriminate between bitter stimuli. *Science* **291**, 1557–1560 (2001).
- Farbman, A. I. Electron microscope study of the developing taste bud in rat fungiform papilla. *Dev. Biol.* **11**, 110–135 (1965).
- Coward, P. *et al.* Controlling signaling with a specifically designed Gi-coupled receptor. *Proc. Natl Acad. Sci. USA* **95**, 352–357 (1998).
- Glendinning, J. I., Gresack, J. & Spector, A. C. A high-throughput screening procedure for identifying mice with aberrant taste and oromotor function. *Chem. Senses* **27**, 461–474 (2002).
- McBride, K., Slotnick, B. & Margolis, F. L. Does intranasal application of zinc sulfate produce anosmia in the mouse? An olfactometric and anatomical study. *Chem. Senses* **28**, 659–670 (2003).
- Franchini, L. F., Rubinstein, M. & Vivas, L. Reduced sodium appetite and increased oxytocin gene expression in mutant mice lacking beta-endorphin. *Neuroscience* **121**, 875–881 (2003).
- Nelson, G. *et al.* Mammalian sweet taste receptors. *Cell* **106**, 381–390 (2001).
- Dahl, M., Erickson, R. P. & Simon, S. A. Neural responses to bitter compounds in rats. *Brain Res.* **756**, 22–34 (1997).
- Nelson, G. *et al.* An amino-acid taste receptor. *Nature* **416**, 199–202 (2002).
- Danilova, V. & Hellekant, G. Comparison of the responses of the chorda tympani and glossopharyngeal nerves to taste stimuli in C57BL/6J mice. *BMC Neurosci.* **4**, 5 (2003).
- Nakashima, K. & Ninomiya, Y. Transduction for sweet taste of saccharin may involve both inositol 1,4,5-trisphosphate and cAMP pathways in the fungiform taste buds in C57BL mice. *Cell. Physiol. Biochem.* **9**, 90–98 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. Liu for help generating T2R5 knockout animals. We are also grateful to A. Cho and S. Taduru for generation of transgenic lines, and to A. Leslie for preparation of antibodies. We thank members of the Zuker and Ryba laboratories for valuable comments and advice. This work was supported in part by a grant from the National Institute on Deafness and Other Communication Disorders to C.S.Z. C.S.Z. is an investigator of the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.S.Z. (charles@flyeye.ucsd.edu) or N.J.P.R. (nick.ryba@nih.gov).

Spatial bistability of Dpp–receptor interactions during *Drosophila* dorsal–ventral patterning

Yu-Chiun Wang¹ & Edwin L. Ferguson^{1,2}

¹Department of Organismal Biology and Anatomy and ²Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, Illinois 60637, USA

In many developmental contexts, a locally produced morphogen specifies positional information by forming a concentration gradient over a field of cells¹. However, during embryonic dorsal–ventral patterning in *Drosophila*, two members of the bone morphogenetic protein (BMP) family, Decapentaplegic (Dpp) and Screw (Scw), are broadly transcribed but promote receptor-mediated signalling in a restricted subset of expressing

cells^{2–4}. Here we use a novel immunostaining protocol to visualize receptor-bound BMPs and show that both proteins become localized to a sharp stripe of dorsal cells. We demonstrate that proper BMP localization involves two distinct processes. First, Dpp undergoes directed, long-range extracellular transport. Scw also undergoes long-range movement, but can do so independently of Dpp transport. Second, an intracellular positive feedback circuit promotes future ligand binding as a function of previous signalling strength. These data elicit a model in which extracellular Dpp transport initially creates a shallow gradient of BMP binding that is acted on by positive intracellular feedback to produce two stable states of BMP–receptor interactions, a spatial bistability in which BMP binding and signalling capabilities are high in dorsal-most cells and low in lateral cells.

The combined activities of both *dpp* and *scw* are necessary for phosphorylation of the BMP signal transducer Mad and ultimately for the specification of dorsal tissues in the *Drosophila* embryo^{3–5}. Although *dpp* is expressed uniformly over the dorsal 40% of the embryonic circumference and *scw* is ubiquitously expressed, at the onset of gastrulation (stage 6) phosphorylated Mad (pMad) staining is restricted to the dorsal 10% of cells that comprise the extraembryonic amnioserosa^{2–4} (Fig. 1c). However, the pattern of pMad staining is dynamic during development. pMad staining is initially visualized during early and mid-stage 5 as a broad, shallow gradient centred on the dorsal midline (Fig. 1a, b). Over the 30 min period between mid-stage 5 and stage 6, the intensity of pMad staining refines, decreasing laterally but significantly increasing dorsally^{6,7} (Fig. 1d).

Because this pattern of BMP signalling does not result from restricted expression of BMP receptors^{8–11}, we determined whether it reflects region-specific ligand–receptor interaction by developing a technique, called perivitelline injection (PVI), to visualize secreted, receptor-bound Dpp. We injected anti-green fluorescent protein (GFP) antibody into the embryonic perivitelline space, the extracellular space between the cell membrane and the vitelline membrane, to detect a biologically active, GFP-tagged form¹² of Dpp expressed under control of the *even-skipped* stripe 2 (*eve-st2*) enhancer¹³ (Fig. 1e). Because injected antibody does not have access to the cytoplasm, it can only interact with secreted Dpp–GFP. After fixation and removal of the vitelline membrane, only antibody bound to Dpp–GFP that is associated with the embryo will be visualized.

Using PVI, we observed that Dpp–GFP is localized to a stripe on the dorsal side of the embryo in a pattern distinctly different from that obtained by conventional immunostaining, which visualizes Dpp–GFP at its site of production, probably in the secretory pathway (Fig. 1f–h). PVI initially detects low levels of Dpp–GFP at mid-stage 5 in a broad dorsal domain centred on the *eve-st2* region (Fig. 1g). By early stage 6, a narrow stripe of Dpp–GFP, which extends over the entire anterior–posterior length of the embryo (Fig. 1h), closely mirrors the spatial extent of pMad staining (Fig. 1j). The immunofluorescence appears to be associated with cell membranes and in punctate structures within cells (Fig. 1i), representing Dpp–GFP bound to the cell surface or internalized, presumably in association with its receptor. The ability to detect the dorsal stripe of Dpp–GFP with conventional immunostaining after PVI (Fig. 1k) suggests that PVI stabilizes an otherwise transient process of ligand–receptor interaction and endocytic processing. Using PVI, we confirmed that an HA-tagged form of Dpp expressed in its endogenous domain also localizes dorsally (Fig. 1l). These findings demonstrate that Dpp undergoes long-range, directional extracellular transport to become localized to the dorsal-most cells.

Two secreted proteins, the ventro-laterally expressed Short gastrulation (Sog) and the dorsally expressed Twisted gastrulation (Tsg), form a tripartite complex with Dpp that blocks Dpp–receptor

interactions *in vitro*⁷. Because the tripartite complex has also been hypothesized to mediate Dpp extracellular transport *in vivo*^{14–16}, we used PVI to examine Dpp–GFP localization in *sog* and *tsg* mutants. In embryos heterozygous for *sog*, *tsg* or both genes, we observed a wider dorsal localization of receptor-bound Dpp–GFP (Fig. 2a–c), suggesting, unlike a previous report¹⁵, that both Sog and Tsg are dose-sensitive components of the system for ligand transport. In embryos completely lacking Sog, Dpp–GFP localization was restricted to an area near the Dpp-expressing cells (Fig. 2d). Although a small amount of Sog-independent Dpp diffusion does occur (see Supplementary Fig. 1), the pattern of Dpp–GFP localization clearly indicates that Dpp movement in the absence of Sog is restricted, probably as a result of rapid receptor association and internalization. The direct visualization of Dpp localization supports and extends the previous demonstration of local induction of Dpp target gene expression in a *sog* background¹⁵, and together with this data confirms the central feature of the model^{14–18} that Sog prevents Dpp from binding to local receptors to allow for long-range Dpp transport.

In *tsg* mutants, little, if any, Dpp–GFP in the *eve-st2* domain was detected by PVI (Fig. 2e), indicating that Tsg promotes Dpp–receptor interactions in addition to its previously described Sog-dependent role in Dpp transport¹⁵. To determine whether this second function of Tsg is independent of Sog, we examined localization of Dpp–GFP in *tsg sog* double mutants and found that Dpp–GFP binds receptors at a level in between that for *sog* and *tsg* single mutants (Fig. 2f). The higher level of ligand binding in the double mutant compared with the *tsg* mutant suggests that Sog can form a stable inhibitory complex¹⁹ with Dpp in the absence of Tsg. Conversely, the lower level of ligand binding in the double mutant compared with the *sog* mutant suggests that Tsg promotes Dpp–receptor interaction independently of Sog. Additional observations (see Supplementary Fig. 2) confirm that *tsg* and *tsg sog* embryos have lower levels of BMP signalling than *sog* embryos. We propose that Tsg either assists in the presentation of Dpp to its receptor or antagonizes an unidentified extracellular inhibitor²⁰.

The dorsally expressed metalloprotease Tolloid (Tld) cleaves Sog in the tripartite Dpp/Sog/Tsg complex to facilitate the release of

Dpp from the complex, thereby promoting Dpp–receptor interactions¹⁶. Consistent with this observation, use of PVI showed that there is no receptor-bound Dpp–GFP in *tld* mutants (Fig. 3a). These data support previous models^{14–16} proposing a mechanism for dorsally directed Dpp transport. In these models, ventrally expressed Sog diffuses dorsally to bind to Dpp, resulting in a transient ventral-to-dorsal gradient of the tripartite complex. Reiterated cycles of complex formation, diffusion, and destruction by Tld lead to a net increase in dorsal Dpp ligand.

The second BMP family member, Scw, is necessary for all BMP signalling in somatic cells before gastrulation⁴. Current data suggest that Dpp and Scw signal through separate type I receptors, Thickveins (Tkv) and Saxophone (Sax) respectively, and that Scw/Sax signalling synergizes with, but is dependent upon, Dpp/Tkv signalling^{21,22}. To better understand the synergy between Dpp and Scw, we first examined the localization of receptor-bound Scw using a maternal *P{nos-Gal4-GCN4.bcd3' UTR}* construct to drive anterior expression of a *UAS-scw-HA* transgene. Although conventional immunostaining detected Scw-HA at its site of expression (Fig. 3c), PVI detected a narrow dorsal stripe of Scw-HA that extended along the entire anterior–posterior axis (Fig. 3d). These results indicate that Scw, like Dpp, undergoes long-range movement in the extracellular space to bind to receptors in the dorsal-most cells.

The finding that Scw and Dpp are internalized in the same spatial domain raised the possibility that internalization of each ligand is dependent on the other. Indeed, we found that Dpp does not bind to its receptors in the absence of Scw (Fig. 3b). For technical reasons, however, we were unable to determine whether Scw localization was dependent on Dpp; nevertheless, the previous observation that Scw depends on Dpp to signal^{21,22} is consistent with this hypothesis. These data suggest that the ligand for BMP signalling might be a complex between Dpp and Scw. BMP ligands are secreted as covalently linked dimers, and the expression patterns of *scw* and *dpp* overlap in the wild-type embryo. Thus, it is possible that the postulated complex could represent a heterodimer between Dpp and Scw. Alternatively, Dpp and Scw homodimers could associate to form a higher order complex. Although previous results had

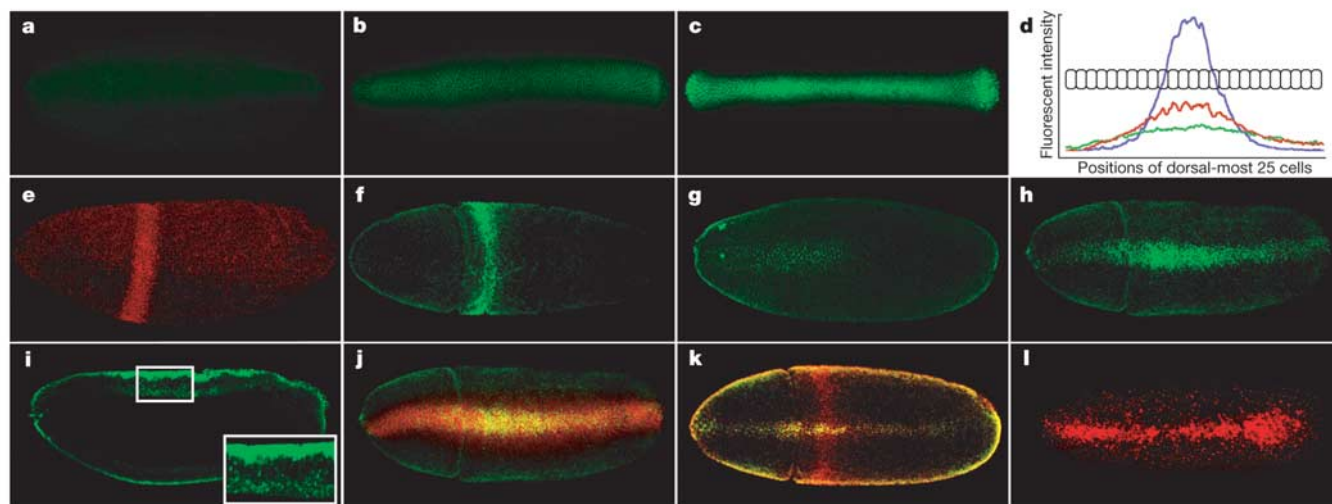


Figure 1 Dpp undergoes long-range, directional transport. **a–c**, pMad staining of wild-type early stage 5 (**a**), mid-stage 5 (**b**), and early stage 6 (**c**) embryos. **d**, Intensity profiles across the dorsal-most 25 cells for pMad staining in **a** (green), **b** (red), and **c** (blue). **e–k**, Embryos carrying two copies of the *P[dpp-GFP.eve.st2]* transgene. **e**, Lateral confocal Z-series projection, *dpp* and *dpp-GFP* transcripts. **f**, Conventional immunostaining using anti-GFP antibody. **g–k**, PVI using anti-GFP antibody. **g**, Mid-stage

5. **h**, Stage 6. **i**, Lateral 10 μ m confocal section. Inset, higher magnification of box. **j**, Double immunostaining of Dpp–GFP (green) and pMad (red). **k**, Double immunostaining with anti-GFP antibody. PVI (green) and conventional method (red). **l**, PVI using an anti-HA antibody of an embryo carrying a Dpp–HA transgene expressed in the endogenous *dpp* domain. All panels except noted are dorsal confocal Z-series projections. All embryos except noted are at stage 6.

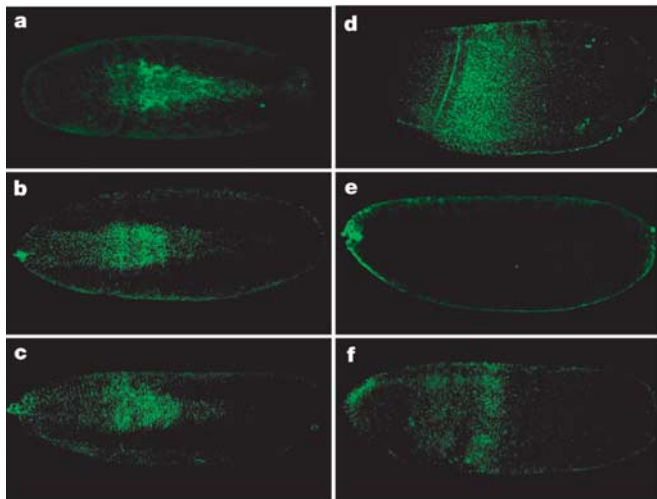


Figure 2 The functions of Sog and Tsg. **a–f**, Confocal Z-series projections of stage 6 embryos carrying two copies of the *P{dpp-GFP.eve.st2}* transgene, processed through PVI of anti-GFP antibody. **a–c**, Dorsal views of Dpp-GFP localization in embryos heterozygous for *sog* (**a**), *tsq* (**b**) or both genes (**c**). **d–f**, Lateral views of Dpp-GFP localization in embryos homozygous for *sog* (**d**), *tsq* (**e**) or both genes (**f**).

suggested that Dpp–Scw heterodimers were not essential for BMP signalling^{21,22}, we revisited this issue by examining pMad staining in double mutant *dpp scw* embryos in which wild-type *dpp* and *scw* were expressed in non-overlapping domains. If obligate heterodimer formation is required for signalling, then such embryos will not exhibit any pMad staining. Conversely, if pMad staining is present, Dpp and Scw homodimers are sufficient for signalling. Although *dpp* mutants carrying the *P{dpp-GFP.eve.st2}* transgene

had a dorsal patch of pMad staining in the *eve-st2* region (see below), no pMad staining was observed in *dpp scw P{dpp-GFP.eve.st2}* embryos (Fig. 3e, f). We injected *scw* messenger RNA either into the posterior-most region or into the *eve-st2* region of these embryos and monitored pMad staining in the injected embryos. Localization of the injected *scw* mRNA was inferred from localization of an equivalent amount of *lacZ* mRNA injected into embryos carrying the *P{dpp-GFP.eve.st2}* transgene (Fig. 3h, j); 85 out of 88 posteriorly injected embryos had clear separation in the spatial extents of *lacZ* and *dpp-GFP* mRNAs. Posterior injection of *scw* mRNA in *dpp scw P{dpp-GFP.eve.st2}* embryos restored pMad staining in the *eve-st2* region (Fig. 3g), indicating that heterodimer formation is not necessary for BMP signalling. Moreover, pMad staining occurred in only the dorsal-most cells, indicating that Sog-mediated transport occurred normally in the injected embryos. The levels of pMad staining in the posteriorly injected embryos were indistinguishable from those in which *scw* mRNA was injected into the *eve-st2* region (Fig. 3i); this indicates that Dpp and Scw homodimers produced in distinct regions of the embryo are sufficient to activate BMP signal transduction to a degree roughly equivalent to that induced in a situation in which Dpp–Scw heterodimers could form. Although these data indicate that Dpp and Scw homodimers are sufficient to activate BMP signalling, they do not exclude the alternative hypothesis that in wild-type embryos, Dpp–Scw heterodimers are formed and participate in BMP signalling²³.

Previous mathematical modelling¹⁵ indicates that competition between the two ligands for the same transport machinery renders the patterning process non-robust. If Sog-mediated transport does not act on Dpp–Scw heterodimers, either Dpp and Scw homodimers must form a complex before their transport by Sog, or Scw must undergo long-range movement that is distinct from Sog-mediated transport of Dpp homodimers. To distinguish between these two alternatives, we asked whether Scw exhibits long-range activity in *sog* mutant embryos. We found that although pMad

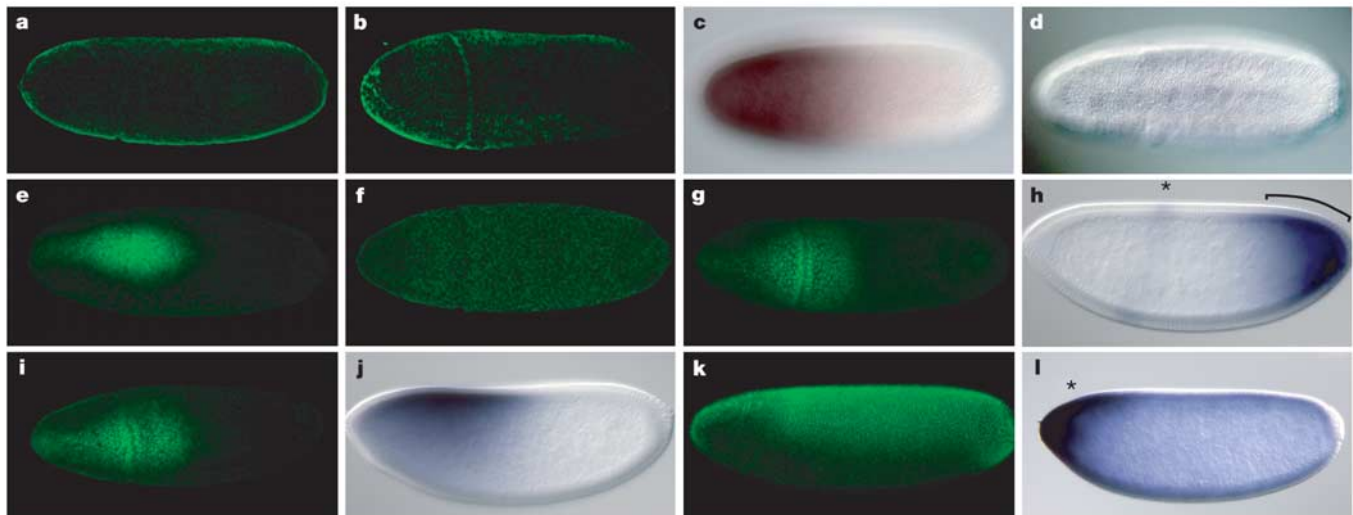


Figure 3 The functions of Tld and Scw. **a, b**, Dorsal confocal Z-series projections of stage 6 *tld* (**a**) or *scw* (**b**) null embryos carrying two copies of the *P{dpp-GFP.eve.st2}* transgene and processed through PVI of anti-GFP antibody. **c, d**, Dorsal Nomarski images of anti-HA immunostaining in embryos containing a *UAS-scw-HA* transgene under the control of a maternal *P[nos-Gal4-GCN4.bcd3' UTR]* driver. **c**, Conventional immunostaining. **d**, PVI. **e–g, i** Dorsal confocal Z-series projections of pMad staining in stage 6 embryos. **e**, *dpp P{dpp-GFP.eve.st2}*. **f, g, i**, *dpp scw P{dpp-GFP.eve.st2}*. **h, j**, Nomarski images of lateral views of stage 5 embryos carrying the *P{dpp-GFP.eve.st2}* transgene hybridized with *gfp* and *lacZ* probes (**h**) or a *lacZ* probe (**j**). **g–j**, Embryos injected with equivalent amount of

scw (**g, i**) or *lacZ* (**h, j**) mRNAs either at the posterior pole (**g, h**) or in the *eve-st2* region (**i, j**). **h**, Injected *lacZ* mRNA (bracket) localized in a region separate from the expression of *dpp-GFP* mRNA (asterisk) from *P{dpp-GFP.eve.st2}*. **j**, Injected *lacZ* mRNA overlaps the *eve-st2* region. **k, l**, *sog; scw* embryos laid by females carrying a *P[scw-bcd]* transgene. **k**, Lateral confocal Z-series projection of pMad staining in an early stage 6 embryo. **l**, Lateral Nomarski image of an early stage 5 embryo hybridized with a *scw* probe. Anterior-localized *scw-bcd* 3' UTR transcripts (asterisk) are superimposed on endogenous mutant *scw* transcripts.

staining was not observed in *sog*; *scw* double mutant embryos (see Supplementary Fig. 3a), anteriorly localized *scw* mRNA expressed from a maternal *P{scw-bcd}* transgene in *sog*; *scw* embryos (Fig. 3l and Supplementary Fig. 3b, c) restored pMad staining to the dorsal 40% of embryonic nuclei along the entire anterior–posterior axis (Fig. 3k), similar to the pattern of pMad staining in *sog* mutants. Thus, localized *scw* expression in *sog*; *scw* mutants results in uniform pMad staining in all Dpp-expressing cells in the embryo, but does not cause pMad staining in any non-Dpp expressing cell, consistent with previous findings that Dpp signals locally in *sog* embryos and that Scw absolutely requires Dpp for signalling^{21,22}. We conclude that a sufficient amount of Scw, but not Dpp (ref. 15 and Fig. 2d), can diffuse in a Sog-independent fashion to promote signalling throughout the embryo, indicating that long-range action of Scw can be independent of Sog. We propose that in the wild type, Scw diffuses freely in the perivitelline space and forms a complex with Dpp homodimers after transport by Sog, but before receptor-mediated endocytosis.

Although many genes necessary for extracellular transport of Dpp are asymmetrically expressed along the dorsal–ventral axis, it is not clear whether their spatial asymmetries are necessary for dorsal transport of Dpp. The one exception is *tsg*, the expression pattern of which is not relevant for normal patterning²⁴. We thus monitored BMP signalling in embryos in which a given gene was expressed in a pattern orthogonal to its normal expression domain. We found that dorsal expression of *tld* was required autonomously for Dpp–receptor interactions, but that ventral expression of *tld* was not sufficient to result in ventral localization of Dpp (see Supplementary Fig. 4). Conversely, although expression of Sog is necessary in order to concentrate Dpp at a distance from the site of its expression^{14,18}, the ventro-lateral expression pattern of *sog* was not obligatory to promote Dpp–receptor interactions dorsally (see Supplementary Fig. 4). Finally, we found that dorsal *dpp* expression also autonomously promotes BMP signalling (see below). Thus, dorsal transport of Dpp is contingent upon the asymmetric expression pattern of at least two genes involved in the transport process.

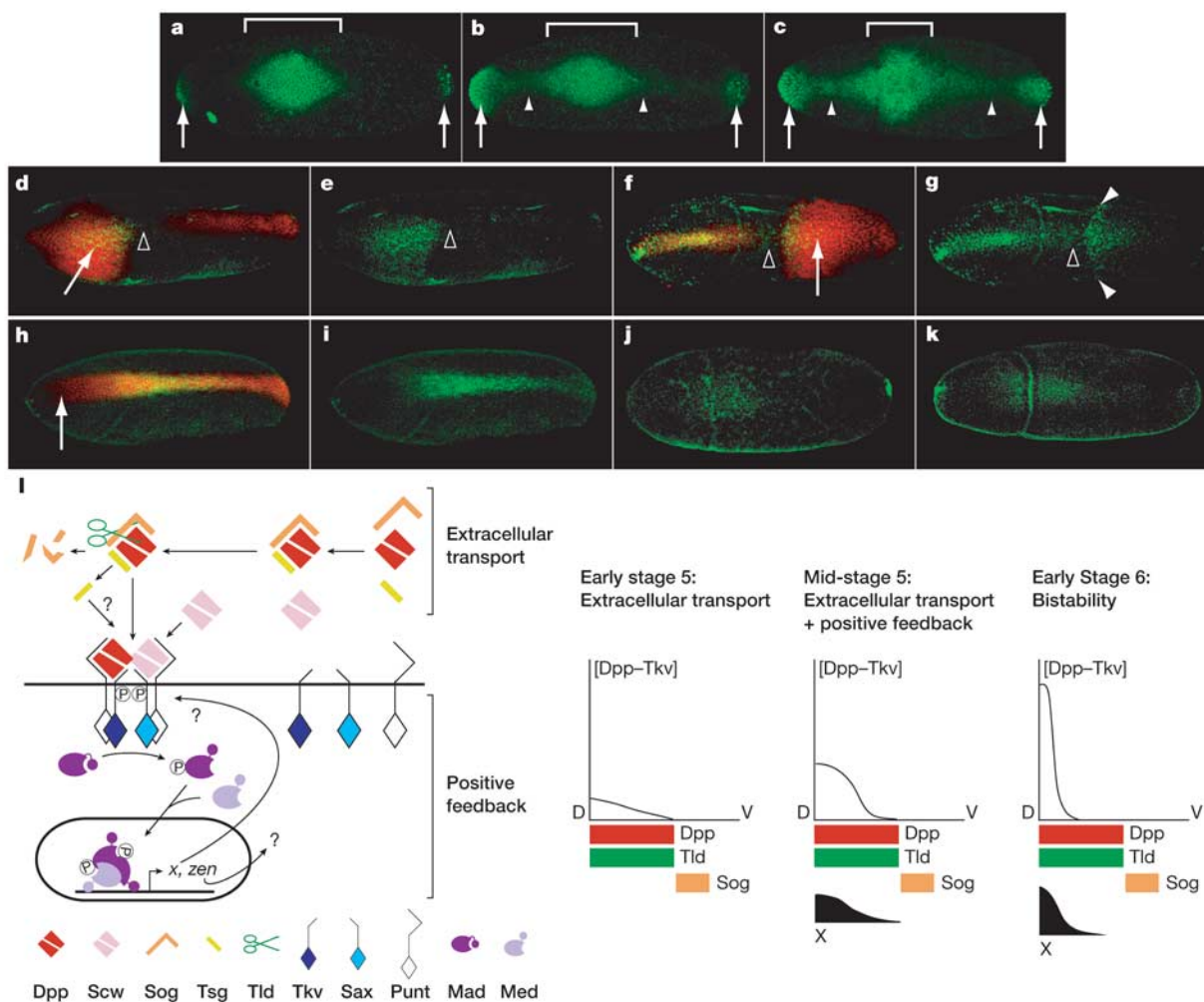


Figure 4 Positive intracellular feedback on Dpp localization and signalling. **a–k**, Dorsal confocal Z-series projections of stage 6 embryos, unless otherwise noted. **a–c**, pMad staining in late stage 5 (**a**), early stage 6 (**b**), and late stage 6 (**c**) *dpp* null mutant embryos carrying the *P{dpp.eve.st2}* transgene. Arrows indicate poles of the embryos; bracket shows *eve-st2* region; arrowheads show spreading of pMad staining outside *eve-st2* region. **d–k**, Embryos carrying either 2 copies (**d–i**, **k**) or 1 copy (**j**) of the *P{dpp-GFP.eve.st2}* transgene. **d–i**, Double immunostaining (PVI+pMad) of embryos injected with *tkv-a* (**d–g**) or *tkv+* (**h**, **i**) mRNA shown in single (**e**, **g**, **i**) or double channels (**d**, **f**, **h**). Dpp–GFP, green; pMad, red. Arrows, sites of mRNA injection; open arrowheads,

reduction of pMad and Dpp–GFP adjacent to *tkv-a* mRNA injection; closed arrowheads, an anterior facing wavefront of Dpp–GFP in posteriorly injected *tkv-a* mRNA embryos. **j**, **k**, PVI using an anti-GFP antibody of embryos completely lacking maternal and zygotic *Medea* activity (**j**), and *zerknüllt* mutant embryos (**k**). **l**, Schematic representation of extracellular transport and intracellular positive feedback and a temporal model for formation of a bistable pattern of Dpp–receptor interactions. 'x' represents one or more transcriptional targets of BMP signalling, the activity of which modulates future Dpp–receptor interactions. See text for details.

To examine the effects of *dpp* expression on BMP signalling, we documented pMad staining in *dpp* null embryos in which *dpp* was expressed in the *eve-st2* pattern. By late stage 5, pMad was visualized in a wide dorsal patch within the *eve-st2* region and at the poles of the embryo (Fig. 4a). Because the *P{dpp.eve.st2}* transgene is the sole source of Dpp expression in these embryos, the polar staining of pMad must arise from non-autonomous, long-range action of Dpp, whereas the lack of pMad staining in the non-polar, non-*eve-st2* nuclei indicates that these regions are less responsive to non-autonomous Dpp ligand. Thus, it seems that there exists an autonomous positive feedback mechanism in non-polar nuclei so that previous BMP signalling promotes future BMP signalling (this feedback mechanism, however, is not due to an effect of BMP signalling on *tld* transcription; see Supplementary Fig. 5). During stage 6, pMad staining gradually appeared in the non-polar, non-*eve-st2* cells (Fig. 4b, c), indicating that continuing accumulation of dorsally transported Dpp partially restores responsiveness to BMP signalling.

To determine whether the positive feedback circuitry promotes future BMP signalling by enhancing the recruitment of transported Dpp, we asked whether injection of mRNAs encoding a constitutively active form²¹ of Tkv (Tkv-a) could lead to ectopic recruitment of Dpp ligand. The mRNA was injected dorsally into syncytial blastoderm (stage 1 or 2) embryos carrying the *P{dpp-GFP.eve.st2}* transgene, and the localization of Dpp-GFP was analysed by PVI at stage 6. When *tkv-a* mRNA was injected anteriorly, Dpp-GFP was observed in a broad patch of cells at the site of injection, as monitored by ectopic pMad staining (Fig. 4d, e). When *tkv-a* mRNA was injected posteriorly, a wavefront of Dpp-GFP localization could be observed at the anterior border of the injected region (Fig. 4f, g). Moreover, in both classes of injected embryos, Dpp-GFP localization at the site of injection was accompanied by a decrease in both Dpp-GFP and pMad in adjacent cells, indicative of a global redistribution of both tagged and endogenous ligands. None of these effects were observed after injection of equivalent amounts of wild-type *tkv* mRNA. Instead, there was either no, or at most a slight, reduction in receptor-bound Dpp-GFP and pMad at the site of injection (Fig. 4h, i). We conclude that previous activation of BMP signalling enhances future interaction between Dpp and its receptor. Although this positive feedback circuit may initially be involved in specifying the dorsal 40% of embryonic cells as competent for accumulation of BMP ligands, it must modulate the spatial extent of ligand binding within the 40% domain, as evidenced both by the redistribution of ligand after *tkv-a* mRNA injection and by the broadened pattern of ligand distribution in Dpp signalling mutants (see below).

We next examined whether the positive feedback circuitry acted through transcriptional targets of BMP signalling. Specifically, we used PVI to examine Dpp-GFP localization in embryos lacking both maternal and zygotic activity of *Medea* (which encodes a Smad4 homologue that forms a complex with activated Mad to regulate BMP-dependent transcription^{6,25}) and also in embryos mutant for the immediate Dpp target gene⁶ *zerknüllt* (which encodes a homeodomain-containing protein²⁶). Embryos of both genotypes exhibit a wider and weaker dorsal localization of receptor-bound Dpp-GFP than do wild-type embryos (Fig. 4j, k), similar to the spatial extent of pMad staining in mid-stage 5 embryos. We argue that positive feedback through BMP-dependent transcription is necessary to refine and intensify Dpp-receptor interactions and that the extracellular transport process alone is not sufficient to produce the narrow stripe of Dpp localization and BMP signalling (Fig. 1a–d) that occurs during normal development.

Experimental and theoretical analyses in other systems have demonstrated that a positive feedback circuit within a signalling pathway can convert a monotonic graded input into either of two stable output states^{27,28}. We propose that BMP signalling during

dorsal-ventral patterning is an example of such a bistable system, but one in which the two stable outputs are manifested in different spatial domains. In our model (Fig. 4l), all embryonic regions are initially able to bind Dpp and signal. Beginning at stage 5, the extracellular transport system creates a transient, shallow gradient of Dpp-receptor interactions with a high point at the dorsal midline (Fig. 1a, b, d) that resembles the computationally simulated profile for Dpp distribution based on a Sog/Tsg-mediated ligand transport model¹⁵. One or more transcriptional targets of this initial graded BMP signalling then positively feeds back to modulate the future capability to bind transported Dpp, either by increasing the affinity of Dpp for its receptor in regions of high signalling (possibly by transcription of a co-receptor), or by decreasing the ability of Dpp to bind to its receptor in regions of low signalling (possibly by promoting post-transcriptional receptor downregulation). This differential Dpp binding capacity causes further dorsal concentration of transported Dpp, resulting by stage 6 in the emergence of a bistable pattern of receptor-ligand interactions (Fig. 1c, d), in which dorsal-most cells retain the ability to bind ligand and to signal while lateral cells have lost these abilities. We speculate that in the absence of the activities of one or more BMP target genes, the difference in Dpp binding capacity between dorsal and lateral cells is diminished, resulting in the observed wider pattern of Dpp localization. Future work will elucidate the mechanism underlying the positive feedback circuit.

Note added in proof: It has come to our attention in two recent papers^{30,31} that the zebrafish homologue of Twisted gastrulation promotes BMP signalling by a mechanism that at least in part is independent of the Sog homologue, Chordin. □

Methods

Drosophila stocks

Flies containing a *P{dpp-GFP.eve.st2}* transgene were generated using standard techniques. A single P element insertion on chromosome II was crossed into stocks containing the following mutations: *sog*^{Y506}, *tsg*^{XB86}, *tsg*^{XB86} *sog*^{Y506}, *tld*^{B4}, *scw*^{s12}, *dpp*^{H46}, *dpp*^{H46} *scw*^{s12}, *Med*¹³ or *zen*⁷. Mutant embryos were identified by lack of β-galactosidase immunostaining of *ftz-lacZ* or *hb-lacZ* transgenes on FM7, CyO, or TM3 balancers. Embryos lacking maternal and zygotic *Med* activity were obtained from germ line clones induced in heterozygous *Med*¹³ females that were mated to *Med*¹³ heterozygous males carrying the Dpp-GFP transgene. Other transgenic flies used were: *P{genomic-Dpp-HA}* and *P{UAS-Scw-HA}* (unpublished gifts from M. O'Connor), *P{UAS-tld.T:HA1}*, *P{sog.eve.st2}*, *P{dpp.eve.st2}*, *P{nos-Gal4-GCN4.bcd3' UTR}*, and *P{scw-bcd}*. All fly stocks are described at Flybase (<http://flybase.org>) unless otherwise noted.

Perivitelline injection of antibody

Embryos were collected, dechorionated, desiccated, mounted with heptane glue on a coverslip, covered under halocarbon oil (700 and 27, 3:1 ratio, Sigma) and aged from stage 4 to early stage 5 depending on the experiment. Approximately 50 picolitres of 1 × PBS-diluted antibody solution was injected into the perivitelline space on the dorsal side²⁹. Injected embryos were incubated at 25 °C for 30–90 min, the vitelline membrane was punctured in a second location, oil was removed by lightly washing with heptane, and embryos on the coverslip were fixed in 3.7% formaldehyde for 18 min. After fixation, the vitelline membrane was removed mechanically, which washes away both free and Dpp-bound antibody in the perivitelline space. The embryos were processed for immunofluorescence staining using standard protocols. Similar staining patterns were obtained whether the antibody was injected into the dorsal or ventral perivitelline space. No staining was detected in embryos that did not express Dpp-GFP.

mRNA injection, immunostaining and *in situ* hybridization

Injection of *tkv*, *scw* and *lacZ* mRNAs into syncytial blastoderm (stage 1 or 2) embryos was performed as described²¹. Approximately 71 picolitres of *tkv-a* (0.73 μg μl⁻¹) or *tkv*⁺ (0.75 μg μl⁻¹) mRNA solution was deposited dorsally, and 21 picolitres of either *scw* (1.54 μg μl⁻¹) or *lacZ* (1.34 μg μl⁻¹) mRNA solution was deposited either at the posterior pole or into the *eve-st2* region. After *tkv* mRNA injection, embryos were aged to between stage 4 and early stage 5, and the PVI protocol was initiated. After *scw* or *lacZ* mRNA injection, embryos were aged to between stage 5 and early stage 6 for subsequent *in situ* hybridization or pMad staining. All immunostainings and *in situ* hybridizations were performed using standard procedures. Fluorescent images were Z-series projections scanned by a Zeiss LSM 510 confocal microscope. Nomarski images were obtained using a Zeiss AxioCam. All images were processed with Adobe Photoshop and figures were compiled using Adobe Illustrator. Additional materials and methods are described in the Supplementary Information.

Received 13 September 2004; accepted 4 January 2005; doi:10.1038/nature03318.

1. Tabata, T. & Takei, Y. Morphogens, their identification and regulation. *Development* **131**, 703–712 (2004).
2. St Johnston, R. D. & Gelbart, W. M. *Decapentaplegic* transcripts are localized along the dorsal-ventral axis of the *Drosophila* embryo. *EMBO J.* **6**, 2785–2791 (1987).
3. Arora, K., Levine, M. S. & O'Connor, M. B. The screw gene encodes a ubiquitously expressed member of the TGF- β family required for specification of dorsal cell fates in the *Drosophila* embryo. *Genes Dev.* **8**, 2588–2601 (1994).
4. Dorfman, R. & Shilo, B. Z. Biphasic activation of the BMP pathway patterns the *Drosophila* embryonic dorsal region. *Development* **128**, 965–972 (2001).
5. Irish, V. F. & Gelbart, W. M. The *decapentaplegic* gene is required for dorsal-ventral patterning of the *Drosophila* embryo. *Genes Dev.* **1**, 868–879 (1987).
6. Rushlow, C., Colosimo, P. E., Lin, M. C., Xu, M. & Kirov, N. Transcriptional regulation of the *Drosophila* gene *zen* by competing Smad and Brinker inputs. *Genes Dev.* **15**, 340–351 (2001).
7. Ross, J. J. *et al.* Twisted gastrulation is a conserved extracellular BMP antagonist. *Nature* **410**, 479–483 (2001).
8. Ferguson, E. L. & Anderson, K. V. Decapentaplegic acts as a morphogen to organize dorsal-ventral pattern in the *Drosophila* embryo. *Cell* **71**, 451–461 (1992).
9. Brummel, T. J. *et al.* Characterization and relationship of Dpp receptors encoded by the *saxophone* and *thick veins* genes in *Drosophila*. *Cell* **78**, 251–261 (1994).
10. Penton, A. *et al.* Identification of two bone morphogenetic protein type I receptors in *Drosophila* and evidence that Brk25D is a decapentaplegic receptor. *Cell* **78**, 239–250 (1994).
11. Letsou, A. *et al.* *Drosophila* Dpp signaling is mediated by the *punt* gene product: a dual ligand-binding type II receptor of the TGF- β receptor family. *Cell* **80**, 899–908 (1995).
12. Teleman, A. A. & Cohen, S. M. Dpp gradient formation in the *Drosophila* wing imaginal disc. *Cell* **103**, 971–980 (2000).
13. Kosman, D. & Small, S. Concentration-dependent patterning by an ectopic expression domain of the *Drosophila* gap gene *knirps*. *Development* **124**, 1343–1354 (1997).
14. Decotto, E. & Ferguson, E. L. A positive role for Short gastrulation in modulating BMP signaling during dorsoventral patterning in the *Drosophila* embryo. *Development* **128**, 3831–3841 (2001).
15. Eldar, A. *et al.* Robustness of the BMP morphogen gradient in *Drosophila* embryonic patterning. *Nature* **419**, 304–308 (2002).
16. Shimmi, O. & O'Connor, M. B. Physical properties of Tld, Sog, Tsg and Dpp protein interactions are predicted to help create a sharp boundary in Bmp signals during dorsoventral patterning of the *Drosophila* embryo. *Development* **130**, 4673–4682 (2003).
17. Holley, S. A. *et al.* The *Xenopus* dorsalizing factor Noggin ventralizes *Drosophila* embryos by preventing DPP from activating its receptor. *Cell* **86**, 607–617 (1996).
18. Ashe, H. L. & Levine, M. Local inhibition and long-range enhancement of Dpp signal transduction by Sog. *Nature* **398**, 427–431 (1999).
19. Marques, G. *et al.* Production of a DPP activity gradient in the early *Drosophila* embryo through the opposing actions of the Sog and Tld proteins. *Cell* **91**, 417–426 (1997).
20. Oelgeschlager, M. *et al.* The pro-BMP activity of Twisted gastrulation is independent of BMP binding. *Development* **130**, 4047–4056 (2003).
21. Neul, J. L. & Ferguson, E. L. Spatially restricted activation of the SAX receptor by SCW modulates DPP/TKV signaling in *Drosophila* dorsal-ventral patterning. *Cell* **95**, 483–494 (1998).
22. Nguyen, M., Park, S., Marques, G. & Arora, K. Interpretation of a BMP activity gradient in *Drosophila* embryos depends on synergistic signaling by two type I receptors, SAX and TKV. *Cell* **95**, 495–506 (1998).
23. Shimmi, O., Umulis, D., Othmer, H. & O'Connor, M. B. Facilitated transport of a Dpp/Scw heterodimer by Sog/Tsg leads to robust patterning of the *Drosophila* blastoderm embryo. *Cell* (in press).
24. Mason, E. D., Williams, S., Grotendorst, G. R. & Marsh, J. L. Combinatorial signaling by Twisted Gastrulation and Decapentaplegic. *Mech. Dev.* **64**, 61–75 (1997).
25. Wisotzkey, R. G. *et al.* *Medea* is a *Drosophila* *Smad4* homolog that is differentially required to potentiate DPP responses. *Development* **125**, 1433–1445 (1998).
26. Rushlow, C., Doyle, H., Hoey, T. & Levine, M. Molecular characterization of the *zerknüllt* gene of the *Antennapedia* gene complex in *Drosophila*. *Genes Dev.* **1**, 1268–1279 (1987).
27. Ferrell, J. E. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr. Opin. Cell Biol.* **14**, 140–148 (2002).
28. Lai, K., Robertson, M. J. & Schaffer, D. V. The Sonic hedgehog signaling system as a bistable genetic switch. *Biophys. J.* **86**, 2748–2757 (2004).
29. Stein, D., Roth, S., Vogelsang, E. & Nusslein-Volhard, C. The polarity of the dorsoventral axis in the *Drosophila* embryo is defined by an extracellular signal. *Cell* **65**, 725–735 (1991).
30. Little, S. C. & Mullins, M. C. Twisted gastrulation promotes BMP signaling in zebrafish dorsal-ventral axial patterning. *Development* **131**, 5825–5835 (2004).
31. Xie, J. & Fisher, S. Twisted gastrulation enhances BMP signaling through chordin dependent and independent mechanisms. *Development* **132**, 383–391 (2005).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank N. Dostani, M. Levine and M. Markstein for fly stocks; S. Cohen, M. Levine and M. Markstein for plasmids; P. ten Dijke and C. Heldin for antibodies; V. Bindokas for assistance with confocal microscopy; S. Lemke and M. Markstein for discussions; D. Bishop, M. O. Casanueva, C. Li, A. Mahowald, J. Malamy, M. Markstein, V. Prince and J. Staley for helpful comments on the manuscript, and M. O'Connor for sharing unpublished data and reagents. Preliminary data on the second function of Tsg were obtained by E. Decotto. Funding was provided by grants to E.L.F. from the NIH and from the Human Frontiers Science program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed E.L.F. (elfergus@midway.uchicago.edu).

The immunoglobulin superfamily protein Izumo is required for sperm to fuse with eggs

Naokazu Inoue¹, Masahito Ikawa², Ayako Isotani^{1,3} & Masaru Okabe^{1,2,3}

¹Genome Information Research Center, ²Research Institute for Microbial Diseases, ³Faculty of Pharmaceutical Sciences, Osaka University, Yamadaoka 3-1, Suita, Osaka 565-0871, Japan

Representing the 60 trillion cells that build a human body, a sperm and an egg meet, recognize each other, and fuse to form a new generation of life. The factors involved in this important membrane fusion event, fertilization, have been sought for a long time¹. Recently, CD9 on the egg membrane was found to be essential for fusion^{2–4}, but sperm-related fusion factors remain unknown. Here, by using a fusion-inhibiting monoclonal antibody⁵ and gene cloning, we identify a mouse sperm fusion-related antigen and show that the antigen is a novel immunoglobulin superfamily protein. We have termed the gene *Izumo* and produced a gene-disrupted mouse line. *Izumo*^{−/−} mice were healthy but males were sterile. They produced normal-looking sperm that bound to and penetrated the zona pellucida but were incapable of fusing with eggs. Human sperm also contain Izumo and addition of the antibody against human Izumo left the sperm unable to fuse with zona-free hamster eggs.

To identify factors involved in sperm–egg fusion, we used a monoclonal antibody, OBF13, against mouse sperm that specifically inhibits the fusion process⁵. The antigen was identified by separation of the crude extracts from mouse sperm by two-dimensional gel electrophoresis and subsequent immunoblotting with the monoclonal antibody. We termed the antigen ‘Izumo’ after a Japanese shrine dedicated to marriage. The identified spot was analysed by liquid chromatography tandem mass spectrometry (LC–MS/MS), and ten peptides that were 100% identical to a part of the sequence listed in the RIKEN full-length database (National Centre for Biotechnology Information (NCBI) accession number XM_133424) were found. The registered DNA sequence was confirmed by sequencing after polymerase chain reaction with reverse transcription (RT–PCR) with total RNA prepared from the testis. A human homologue was found as an unverified gene in the NCBI database (accession number BC034769). The gene encodes a novel immunoglobulin superfamily (IgSF), type I membrane protein with an extracellular immunoglobulin domain that contains one putative glycosylation site (Fig. 1a, b). Mouse Izumo was shown to be a testis (sperm)-specific 56.4-kDa antigen by western blotting with a polyclonal antibody raised against recombinant mouse Izumo (Fig. 1c). Izumo was also detectable as a 37.2-kDa protein by western blotting of human sperm with anti-human Izumo antibody (Fig. 1d). Izumo was not detectable on the surface of fresh sperm. Coinciding with the fact that mammalian sperm are incapable of fertilizing eggs when ejaculated and that fertilization occurs only after an exocytotic process called the acrosome reaction, both mouse and human Izumo became detectable on sperm surface only after the acrosome reaction (Fig. 1e, f). This would probably be because Izumo is not localized on plasma membrane of fresh spermatozoa but is hidden under plasma membrane and accessible after the acrosome reaction, as occurs with CD46 on mouse sperm⁶.

To address the physiological role of Izumo *in vivo* we generated *Izumo*-deficient mice by homologous recombination. An *Izumo*-targeting construct was designed to replace exons 2–10 with a neomycin-resistant gene (*neo*^r) (Fig. 2a). Both the targeting event in D3 embryonic stem cells and the germline transmission of targeted genes were confirmed by Southern blot analysis (Fig. 2b).



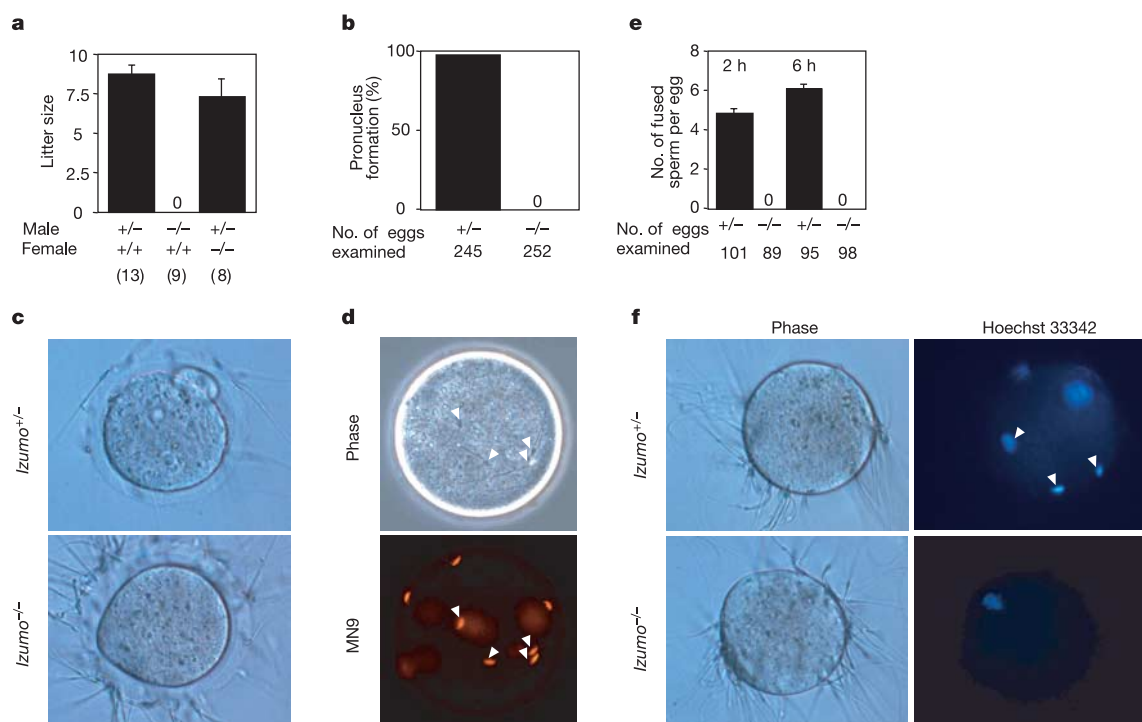


Figure 3 Male infertility caused by *Izumo* disruption. **a**, Fecundity of *Izumo*^{+/+} and *Izumo*^{-/-} males and *Izumo*^{-/-} females. The numbers in parentheses indicate the numbers of mating pairs. **b**, **c**, *In vitro* fertilization of sperm from *Izumo*^{+/+} and *Izumo*^{-/-} mice. Unlike *Izumo*^{+/+}, the eggs inseminated with *Izumo*^{-/-} sperm had many sperm on their zona pellucida, owing to the failure of sperm-egg fusion that probably leads to the absence of zona-reaction to lessen the sperm-binding ability of the zona pellucida. The error bars in **b** are not visible ($n = 5$). **d**, Upper panel, accumulation of

many sperm in the perivitelline space of the eggs recovered from the females mated with *Izumo*^{-/-} males. Lower panel, sperm in perivitelline space labelled with acrosome-reacted, sperm-specific monoclonal antibody MN9 (ref. 15). **e**, Average numbers of fused sperm observed 2 and 6 h after insemination ($n = 5$). **f**, Fused sperm stained by Hoechst 33342 preloaded into the egg. The arrowheads show the fused sperm. Errors bars in **a** and **e** are s.e.m.

Intercrosses between heterozygous F₁ mice yielded offspring that segregated in a mendelian distribution: 43 wild-type, 92 heterozygous and 47 homozygous mutant weaning pups. *Izumo*^{-/-} mutant mice were healthy and showed no overt developmental abnormalities. *Izumo*^{-/-} females demonstrated normal fecundity. *Izumo*^{+/+} males also showed normal fertilizing ability (Fig. 3a). However, *Izumo*^{-/-} males were sterile despite normal mating behaviour and ejaculation, with normal vaginal plug formations. After observation of 28 plugs, nine pairs of *Izumo*^{-/-} male and wild-type females were kept for another 4 months but no pregnancies were observed (Fig. 3a). In at least four different cases of gene knockouts that resulted in male sterility attributed to impaired zona-binding ability, the sperm also failed to migrate into the oviduct^{7,8,11,12}. However, disruption of *Izumo* did not cause any defect in sperm migration into the oviduct (data not shown, and there was no reduction of sperm motility in *Izumo*^{-/-} sperm; motility was measured 120 min after incubation by computer-aided sperm analysis (CASA; mean $\pm$ s.e.m. = $81.7 \pm 7.7\%$ in *Izumo*^{+/+} sperm and $77 \pm 8.9\%$ in *Izumo*^{-/-} sperm)). The sterile nature of *Izumo*^{-/-} sperm was shown in the *in vitro* fertilization system (Fig. 3b, c, and Supplementary Movie 1). The impaired fertilization step undoubtedly followed zona penetration because sperm penetrated the zona pellucida and accumulated in the perivitelline space of the eggs (Fig. 3d).

Syngamy can be considered to occur to two stages: binding of the sperm plasma membrane to that of the egg, and actual membrane fusion. *Izumo*^{-/-} sperm were capable of binding to the plasma membranes of eggs whose zona pellucida had been mechanically removed¹³ (Fig. 3e, f). In this system, the *Izumo*^{+/+} sperm incubated for 2 and 6 h fused to eggs in approximate ratios of 4.5 and 6 sperm per egg, respectively, but no *Izumo*^{-/-} sperm fused with eggs.

Sperm cannot fuse with eggs unless the former have undergone the acrosome reaction¹⁴. To verify the acrosomal status of *Izumo*^{-/-} sperm, we stained the sperm accumulated in perivitelline spaces with the MN9 monoclonal antibody, which immunoreacts only to the equatorial segment of acrosome-reacted sperm¹⁵. The staining indicated that the *Izumo*^{-/-} sperm had undergone the acrosome reaction (Fig. 3d) but failed to fuse with eggs.

Because no offspring were fathered by *Izumo*^{-/-} male mice, it was unclear whether the defect was limited to fusion or extended to later developmental stages. To address this question, we used intracytoplasmic sperm injection (ICSI) to insert *Izumo*^{-/-} sperm directly into the cytoplasm of wild-type eggs and bypass the fusion step¹⁶. Eggs injected with *Izumo*^{-/-} sperm were successfully activated and the fertilized eggs were transplanted into the oviducts of pseudopregnant females. The eggs implanted normally and the resulting embryos developed appropriately to term with rates similar to those of heterozygous mice (Table 1).

Sperm-egg fusion is known to be less species-specific than sperm-zona interaction. For example, human sperm cannot penetrate the hamster zona pellucida but they can fuse with zona-free hamster eggs, and this system (zona-free hamster-egg sperm

Table 1 Development of eggs after ICSI with *Izumo*^{-/-} sperm

Sperm	No. of eggs used	No. of eggs surviving after ICSI	No. of eggs developing to two-cell stage	No. of pups born
<i>Izumo</i> ^{-/-}	95	59	42 (71%)*	12 (29%)+
<i>Izumo</i> ^{+/+}	85	54	43 (80%)	6 (14%)

*Percentages are based on numbers of eggs surviving after ICSI.

+All offspring from *Izumo*^{-/-} sperm were confirmed to possess the *Izumo*-null allele.

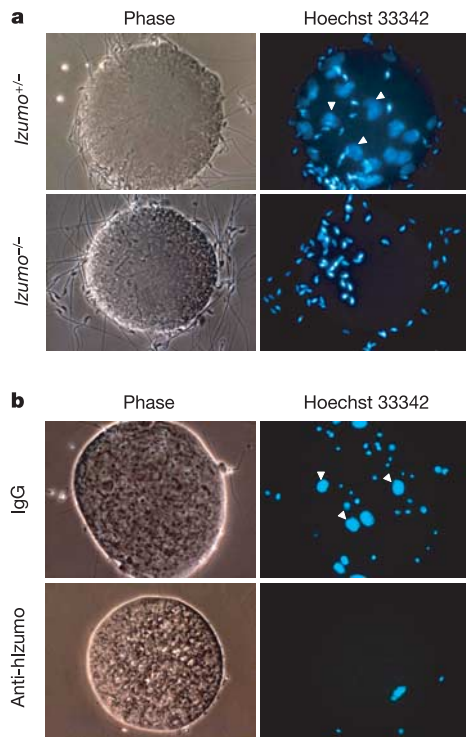


Figure 4 Involvement of Izumo in a xeno-species fusion system. **a**, At 6 h after the insemination of zona-free hamster eggs with *Izumo*^{+/+} and *Izumo*^{-/-} mouse sperm, sperm heads were stained by adding Hoechst 33342 to the medium. The sperm-egg binding was strong enough to resist repeated pipetting. **b**, Human sperm were also added with 25 $\mu\text{g ml}^{-1}$ anti-human Izumo (anti-hIzumo) or control IgG to zona-free hamster eggs ($n=3$). No fusion was observed in the presence of anti-Izumo antibody. Arrowheads indicate the swelling sperm head after staining with Hoechst 33342. The eggs were pressed under a coverslip to bring many sperm into focus.

penetration test) has been used for the assessment of human sperm fertility¹⁷. We first examined the contribution of mouse Izumo in a zona-free hamster-egg sperm penetration assay. As indicated in Fig. 4a, the mouse Izumo was essential not only in the homologous fusion system but also for heterologous fusion with hamster eggs. Similarly, when the anti-human Izumo polyclonal antibody was added to the incubation mixture, no fusion was observed, whereas the sperm treated with control IgG fused with eggs at an average of 5.9 ± 0.7 sperm per egg. The total numbers of eggs observed were 23 and 29, respectively ($n=3$). These results indicated that human Izumo is involved in the fertilization process in human sperm (Fig. 4b). However, further investigation will be required to explain the function of Izumo in human fertilization because adding the antibody caused inhibition of human sperm binding to the egg plasma membrane in the heterologous sperm-egg fusion system.

The phenotypes of gene knockout mice are not always related to the disrupted genes but are sometimes caused by disruption of a neighbouring gene¹⁸. To examine whether the phenotype was directly derived from the lack of Izumo on sperm, we performed a rescue experiment by crossing *Izumo*^{-/-} mice with transgenic mouse lines generated to express Izumo by using the testis-specific calmodulin promoter¹². The sterile phenotype was rescued with the transgenically expressed Izumo on mouse sperm (Supplementary Fig. 1).

In the search for sperm surface proteins that function in sperm-egg plasma-membrane binding and fusion, various candidates such as DE¹⁹, CD46 (ref. 20), equatorin¹⁵, Sperad²¹ and SAMP32 (ref. 22) have been reported. ADAM family proteins are given the most attention for their possession of a putative fusion peptide (ADAM1) and disintegrin domain (ADAM2 and ADAM3)²³. None of the mice

possessing disrupted ADAM1a, ADAM2 and ADAM3 show a significant defect in the ability to fuse with eggs^{7,8,24}, but do show an impairment of sperm-zona binding ability. Similarly, CD46 disruption does not diminish fusion⁶. In contrast, CD9 on the egg surface is essential for the fusing ability of eggs²⁻⁴ and some indications for the involvement of the binding of integrins to CD9 are postulated in reference to sperm-egg fusion. However, the disruptions of the most probable candidate integrins $\alpha 6$ and $\beta 1$ cause no major influence on the fusing ability of eggs²⁵. Thus, for several years, postulated fertilization mechanisms were repeatedly changed as a result of gene disruption experiments. This suggests that the essential nature of the candidate gene must be judged after observing the phenotype of the gene-disrupted mice. In this context, Izumo is the first sperm membrane protein shown to be essential for fusion. It is not yet known whether sperm Izumo interacts with egg CD9, as occurs with placental IgSF protein PSG17 (ref. 26); neither do we know why the localization of Izumo after acrosome reaction is not limited to the equatorial segment where fusion initially takes place. All we can say now is that continued study of this protein's function will undoubtedly lead to a fuller understanding of the cell-cell fusion process in fertilization and perhaps in other somatic systems such as muscle cells or trophoblasts.

The finding not only provides insight into the enigmatic fusion mechanism but also promises benefits in the clinical treatment of infertility and the potential development of new contraceptive strategies. □

Methods

Cloning of Izumo

Izumo amino acid sequences were determined by combining two-dimensional gel electrophoresis and LC-MS/MS. Some peptide sequences (Fig. 1a, shown in red) were analysed and identified as members of an immunoglobulin superfamily protein (NCBI accession number XM_133424) whose function was not clarified. To confirm the DNA sequence we used RT-PCR to amplify Izumo from mouse testis RNA as a template with primers derived from this sequence. The polyclonal antibodies against mouse and human Izumo were produced by immunizing an Izumo-expressing RK13 cell line into rabbits²⁷.

Generation of *Izumo* knockout mice

A targeting vector was constructed with the use of pMulti-ND 1.0 containing the Neo-resistance gene (*neo*^r) as a positive selection marker and diphtheria toxin A chain (DT) as a negative selection marker (provided by J. Takeda and T. Ijiri). A 1.7-kb *AscI*-*PacI* fragment and a 6.7-kb *Clat*-*XhoI* fragment were inserted as a short and long arm, respectively. Embryonic stem cells derived from 129/Sv (D3) were electroporated with *PmeI*-digested linearized DNA. Of 385 G418-resistant clones, four had undergone homologous recombination correctly. Three targeted cell lines were injected into C57BL/6 blastocysts, resulting in the birth of male chimaeric mice. These mice were then crossed with C57BL/6 to obtain heterozygous mutants. Mice used in the study were the offspring of crosses between F₁ and/or F₂ generations.

In vitro fertilization

Mouse sperm were collected from cauda epididymides and capacitated *in vitro* for 2 h in a 200- μl drop of TYH medium²⁸ covered with paraffin oil. Wild-type female mice (more than 8 weeks old) were superovulated by the injection of 5 U of human chorionic gonadotropin (hCG) 48 h after a 5-U injection of pregnant mare serum gonadotropin. The eggs were collected from the oviduct 14 h after the hCG injection. Eggs were placed in a 200- μl drop of TYH medium. These eggs were incubated with 2×10^5 *Izumo*^{+/+} or *Izumo*^{-/-} sperm per ml incubated for 2 h at 37 °C in 5% CO₂, and unbound sperm were washed away. Eggs were observed 6 h after insemination for the formation of pronuclei under a Hoffman modulation contrast microscope.

Zona-free-egg sperm penetration assay

After being freed from cumulus cells with 0.01% (w/v) hyaluronidase, the zona pellucida was removed from mouse or hamster eggs with a piezo-manipulator as described previously¹³. Fusion assessment was performed in two different ways. In the first method, zona-free mouse oocytes were preloaded with Hoechst 33342 by incubating them with the dye (1 $\mu\text{g ml}^{-1}$) in TYH for 10 min and washing them before addition of the sperm. After 30 min of incubation, the eggs were observed under a fluorescence microscope (excitation with ultraviolet) after fixing with 0.25% glutaraldehyde. This procedure enabled the staining of only fused sperm nuclei by transferring the dye into sperm after membrane fusion as in Fig. 3. Alternatively, in the second method the zona-free hamster eggs were incubated for 6 h with sperm without preloading of the dye. By this time, enlarged sperm heads could be seen when fusion occurred. The eggs with sperm were incubated in 1 $\mu\text{g ml}^{-1}$ Hoechst 33342 for 10 min to ensure that all bound (original sperm head shape) and fused (enlarged round shape) sperm would be stained by the dye, as in Fig. 4. In all

experiments the human sperm were collected from liquefied semen by the swim-up method and incubated for 6 h before addition to eggs. We used BWW medium²⁹ containing 35 mg ml⁻¹ human serum albumin (HSA) for the human sperm experiment.

Received 12 October 2004; accepted 17 January 2005; doi:10.1038/nature03362.

- Stein, K. K., Primakoff, P. & Myles, D. Sperm-egg fusion: events at the plasma membrane. *J. Cell Sci.* **117**, 6269–6274 (2004).
- Miyado, K. *et al.* Requirement of CD9 on the egg plasma membrane for fertilization. *Science* **287**, 321–324 (2000).
- Le Naour, F., Rubinstein, E., Jasmin, C., Prenant, M. & Boucheix, C. Severely reduced female fertility in CD9-deficient mice. *Science* **287**, 319–321 (2000).
- Kaji, K. *et al.* The gamete fusion process is defective in eggs of CD9-deficient mice. *Nature Genet.* **24**, 279–282 (2000).
- Okabe, M. *et al.* Capacitation-related changes in antigen distribution on mouse sperm heads and its relation to fertilization rate *in vitro*. *J. Reprod. Immunol.* **11**, 91–100 (1987).
- Inoue, N. *et al.* Disruption of mouse CD46 causes an accelerated spontaneous acrosome reaction in sperm. *Mol. Cell. Biol.* **23**, 2614–2622 (2003).
- Nishimura, H., Kim, E., Nakanishi, T. & Baba, T. Possible function of the ADAM1a/ADAM2 fertilin complex in the appearance of ADAM3 on the sperm surface. *J. Biol. Chem.* **279**, 34957–34962 (2004).
- Cho, C. *et al.* Fertilization defects in sperm from mice lacking fertilin beta. *Science* **281**, 1857–1859 (1998).
- Saxena, D. K., Oh-Oka, T., Kadomatsu, K., Muramatsu, T. & Toshimori, K. Behaviour of a sperm surface transmembrane glycoprotein basigin during epididymal maturation and its role in fertilization in mice. *Reproduction* **123**, 435–444 (2002).
- Bookbinder, L. H., Cheng, A. & Bleil, J. D. Tissue- and species-specific expression of sp56, a mouse sperm fertilization protein. *Science* **269**, 86–89 (1995).
- Hagaman, J. R. *et al.* Angiotensin-converting enzyme and male fertility. *Proc. Natl Acad. Sci. USA* **95**, 2552–2557 (1998).
- Ikawa, M. *et al.* Calmegin is required for fertilin alpha/beta heterodimerization and sperm fertility. *Dev. Biol.* **240**, 254–261 (2001).
- Yamagata, K. *et al.* Sperm from the calmegin-deficient mouse have normal abilities for binding and fusion to the egg plasma membrane. *Dev. Biol.* **250**, 348–357 (2002).
- Yanagimachi, R. *Mammalian Fertilization* (eds Knobil, E. & Neill, J. D.) (Raven, New York, 1994).
- Manandhar, G. & Toshimori, K. Exposure of sperm head equator after acrosome reaction and its fate after fertilization in mice. *Biol. Reprod.* **65**, 1425–1436 (2001).
- Kimura, Y. & Yanagimachi, R. Intracytoplasmic sperm injection in the mouse. *Biol. Reprod.* **52**, 709–720 (1995).
- Yanagimachi, R., Yanagimachi, H. & Rogers, B. J. The use of zona-free animal ova as a test-system for the assessment of the fertilizing capacity of human spermatozoa. *Biol. Reprod.* **15**, 471–476 (1976).
- Olson, E. N., Arnold, H. H., Rigby, P. W. & Wold, B. J. Know your neighbors: three phenotypes in null mutants of the myogenic bHLH gene MRF4. *Cell* **85**, 1–4 (1996).
- Rochwerger, L., Cohen, D. J. & Cusanu, P. S. Mammalian sperm-egg fusion: the rat egg has complementary sites for a sperm protein that mediates gamete fusion. *Dev. Biol.* **153**, 83–90 (1992).
- Anderson, D. J., Abbott, A. F. & Jack, R. M. The role of complement component C3b and its receptors in sperm-oocyte interaction. *Proc. Natl Acad. Sci. USA* **90**, 10051–10055 (1993).
- Ilayperuma, I. Identification of the 48-kDa G11 protein from guinea pig testes as sperad. *J. Exp. Zool.* **293**, 617–623 (2002).
- Hao, Z. *et al.* SAMP32, a testis-specific, isoantigenic sperm acrosomal membrane-associated protein. *Biol. Reprod.* **66**, 735–744 (2002).
- Blöbel, C. P. *et al.* A potential fusion peptide and an integrin ligand domain in a protein active in sperm-egg fusion. *Nature* **356**, 248–252 (1992).
- Nishimura, H., Cho, C., Branciforte, D. R., Myles, D. G. & Primakoff, P. Analysis of loss of adhesive function in sperm lacking cyritestin or fertilin beta. *Dev. Biol.* **233**, 204–213 (2001).
- He, Z. Y. *et al.* None of the integrins known to be present on the mouse egg or to be ADAM receptors are essential for sperm-egg binding and fusion. *Dev. Biol.* **254**, 226–237 (2003).
- Ellerman, D. A., Ha, C., Primakoff, P., Myles, D. G. & Dvorkin, G. S. Direct binding of the ligand PSG17 to CD9 requires a CD9 site essential for sperm-egg fusion. *Mol. Biol. Cell* **14**, 5098–5103 (2003).
- Inoue, N. *et al.* A novel chicken membrane-associated complement regulatory protein: molecular cloning and functional characterization. *J. Immunol.* **166**, 424–431 (2001).
- Toyoda, Y., Yokoyama, M. & Hoshi, T. Studies on the fertilization of mouse egg *in vitro*. *Jpn. J. Anim. Reprod.* **16**, 147–151 (1971).
- Overstreet, J. W., Yanagimachi, R., Katz, D. F., Hayashi, K. & Hanson, F. W. Penetration of human spermatozoa into the human zona pellucida and the zona-free hamster egg: a study of fertile donors and infertile patients. *Fertil. Steril.* **33**, 534–542 (1980).
- Nakanishi, T. *et al.* Real-time observation of acrosomal dispersal from mouse sperm using GFP as a marker protein. *FEBS Lett.* **449**, 277–283 (1999).
- Okabe, M. *et al.* A human sperm antigen possibly involved in binding and/or fusion with zona-free hamster eggs. *Fertil. Steril.* **54**, 1121–1126 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Toshimori for providing anti-MN9 antibody; K. Yamagata for discussions; G. L. Gerton and S. Moss for critically reviewing the draft; and Y. Maruyama, A. Kawai and Y. Koreeda for technical assistance with gene disruption. This work was supported by grant-in-aid for Scientific Research and the 21st Century COE program from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.O. (e-mail: okabe@gen-info.osaka-u.ac.jp). Sequences have been deposited with GenBank under accession numbers AB195681 for mouse *Izumo*, AB195682 for human *Izumo* and AB195683 for rat *Izumo*.

Agonist/endogenous peptide–MHC heterodimers drive T cell activation and sensitivity

Michelle Krogsgaard¹, Qi-jing Li¹, Cenik Sumen^{1*}, Johannes B. Huppa^{1,2}, Morgan Huse¹ & Mark M. Davis^{1,2}

¹The Department of Microbiology and Immunology, and ²Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA

* Present address: Center for Blood Research, Harvard Medical School, Boston, Massachusetts 02115, USA

$\alpha\beta$ T lymphocytes are able to detect even a single peptide–major histocompatibility complex (MHC) on the surface of an antigen-presenting cell^{1,2}. This is despite clear evidence, at least with CD4⁺ T cells, that monomeric ligands are not stimulatory^{3,4}. In an effort to understand how this remarkable sensitivity is achieved, we constructed soluble peptide–MHC heterodimers in which one peptide is an agonist and the other is one of the large number of endogenous peptide–MHCs displayed by presenting cells. We found that some specific combinations of these heterodimers can stimulate specific T cells in a CD4-dependent manner. This activation is severely impaired if the CD4-binding site on the agonist ligand is ablated, but the same mutation on an endogenous ligand has no effect. These data correlate well with analyses of lipid bilayers and cells presenting these ligands, and indicate that the basic unit of helper T cell activation is a heterodimer of agonist peptide– and endogenous peptide–MHC complexes, stabilized by CD4.

It has been known for some time that $\alpha\beta$ T-cell antigen receptor (TCR)-bearing T cells are selected for weak or rare interactions with some of the thousands of self-peptide–MHC (pMHC) complexes (that is, endogenous peptide–MHC complexes) presented in the thymus and also require the presence of these complexes in the periphery to survive^{5–7}. However, the role of self-peptide–MHC complexes in the activation of mature T cells is controversial. Although there is some evidence to suggest that endogenous peptides or their mimics enhance CD4⁺ helper T cell activation⁸, and that T cell responsiveness declines rapidly in the absence of endogenous pMHC contact⁷, others working with CD8⁺ T cells have found no effect⁹. Recently, we have found that there is a substantial (~20%) recruitment of endogenous peptide–MHC molecules into the immunological synapse^{1,8}. This seems to be driven by weak interactions with TCRs. This phenomenon, together with data showing that even a single agonist peptide–MHC complex can initiate T cell activation^{1,2} has led to the suggestion of a ‘pseudodimer’ model of T cell activation, in which agonist and endogenous peptide–MHC complexes, stabilized by CD4, are crucial intermediates for triggering CD4⁺ T lymphocytes¹.

To test this model, we used the well-characterized moth cytochrome *c* (MCC) system in which specific T cells bearing the 5C.C7 TCR recognize MCC bound to the murine MHC class II molecule I-E^k (ref. 10). We looked at the T cell–antigen-presenting cell (APC) synapse and measured the accumulation of MCC and its variants as well as a spectrum of self-peptide ligands previously found to be associated with I-E^k (refs 8, 11, 12; Fig. 1a). In these assays, we used *in vitro*-stimulated lymph node cells isolated from 5C.C7- $\alpha\beta$ TCR transgenic mice¹⁰. These T cell blasts were mixed with APCs that had been loaded with a strong agonist variant of MCC (K5), and the peptide being analysed was labelled with Cy3. The mixture was imaged by time-lapse three-dimensional (3D) fluorescence microscopy. Figure 1b shows representative images of pMHC accumulation at the T cell–APC interface. Consistent with previous

results⁸, we observed strong accumulation of K5–I-E^k, MCC–I-E^k and 99A–I-E^k complexes, but not 99E–I-E^k (Fig. 1b). Endogenous peptides derived from the cysteine protease of the endoplasmic reticulum (ER60) and the complement cytotoxicity inhibitor (cCyt) accumulated at ratios comparable to MCC and K5. Accumulation of 99A, MSA and β -2-microglobulin (β 2M) was slightly weaker, and heat shock protein 70 (HSP70) showed little or no accumulation (Fig. 1a, Supplementary Fig. S1). Some peptides did not show staining levels above background in APCs, and their accumulation could not be assessed. The other peptides bind I-E^k very stably (Supplementary Fig. S2). The peptide ligand Unknown1 shows high levels of accumulation at the synapse (data not shown), but it is not clear whether this is an endogenous ligand. Almost all of the endogenous peptides tested were able to accumulate at the interface, indicating that this phenomenon might be largely driven by TCR–MHC binding.

Despite their accumulation at the synapse, none of the endogenous peptides shown in Fig. 1a have any T cell-stimulatory activity individually (Supplementary Fig. S2). To determine whether they had any co-agonist activity¹³, we made soluble heterodimers⁴ by pairing endogenous peptides with agonist ligands. These dimers were added to 5C.C7 T cell blasts loaded with the calcium indicator dye fura-2 (ref. 14) and imaged by time-lapse, 3D fluorescence microscopy (Supplementary Videos S1–S5). Consistent with previous studies⁴, homodimers of K5–I-E^k (K5–K5) activate these T cells as measured by an increase in free calcium (Fig. 2a). Furthermore, I-E^k-heterodimers that combine K5 with some of the ‘null’ or self-peptides (99A, ER60, MSA and cCyt) can also activate T cells (Fig. 2c). In contrast, heterodimers containing β 2M, HSP70 or 99E do not activate T cells (Fig. 2c). When MCC is used as the agonist, stimulation is weaker (Supplementary Fig. S3c, d) and only the strongest co-agonists (99A and ER60) can induce T-cell activation (Supplementary Fig. S3c, d). These data lend support to recent modelling studies, which suggest that the activity of a self-peptide will be dependent on the strength of the agonist ligand¹⁵.

Some heterodimers could also activate naive T cells, inducing a strong Ca²⁺ response even without co-stimulation (Supplementary Fig. S3e), showing that this effect is not limited to previously activated cells. Figure 2b shows a time course of Ca²⁺ elevation for some of these constructs. We find that K5–K5, K5–ER60 and K5–99A induce a sustained increase in Ca²⁺, with fura-2 ratios remaining roughly 1.5–2-fold above background (Fig. 2b). There is no increase in Ca²⁺ for K5–99E or K5– β 2M. In terms of the percentage of T cells showing a rise in Ca²⁺ levels (Fig. 2c), both K5–99A and K5–ER60 heterodimers are close to K5–K5 homodimers in potency, whereas the MSA and cCyt heterodimers are much weaker.

We also determined the dose-response of T cell activation (Fig. 2d) and analysed the relative Ca²⁺ response after pMHC dimer activation (Supplementary Fig. S4a). Another measure of early T cell activation is the local generation of 3'-phosphoinositides (3'-PIP₃) by phosphatidylinositol-3-OH kinase (PI(3)K)¹⁶ at the T cell membrane. This can be visualized by monitoring the pleckstrin homology (PH) domain of the serine/threonine kinase Akt (also known as protein kinase B) fused to yellow fluorescent protein (YFP), PH(Akt)–YFP¹⁷. As shown in Fig. 2a, PI(3)K activity at the plasma membrane exactly parallels the results obtained with the Ca²⁺ analysis (Fig. 2e, Supplementary Fig. 4b, c and Supplementary Videos S1–S5).

To assess whether dimer stimulation results in complete T cell activation, we used intracellular antibody staining for interleukin-2 (IL-2) (Fig. 3a, b). We find that the same heterodimers that induce the first indicators of T cell activation can also, after four hours, stimulate IL-2 production and its loading into secretory vesicles.

Because the covalently linked dimers used here might not reflect the physiological situation on the cell surface, we also analysed the

contribution of the endogenous peptides to T cell stimulation when presented by membrane-associated MHC. Here we used the MCC peptide (instead of the more potent agonist K5), because its characteristics are more typical of helper T cell antigens. We pulsed Chinese hamster ovary (CHO) cells expressing glycan-phosphatidyl inositol (GPI)-anchored I-E^k molecules¹⁸ with agonist MCC peptide diluted with the various endogenous peptides (Fig. 3c and Supplementary Fig. S5). GPI-linked molecules do not cross the endosomal compartment and thus these MHCs lack bound peptides¹⁹. We also used supported lipid bilayers^{8,20} to present both agonist and endogenous pMHCs. We diluted His-tagged I-E^k (with MCC covalently linked to the I-E^k β -chain) into I-E^k pulsed with the various endogenous peptides, and incorporated the resulting pMHC molecules into Ni-lipid-containing planar bilayers²⁰ (Fig. 3d and Supplementary Fig. S5). In accordance with the pMHC dimer experiments, we find that low densities of agonist can synergize with the self-peptide ER60 to activate the T cells (Fig. 3c, d), but this synergy does not occur with β 2M. We also find that MSA and cCyt can contribute to activation by MCC, although significantly more weakly than ER60 (Supplementary Fig. S5). This is in line with the soluble K5 heterodimer data, but in contrast with the MCC heterodimers, for which MSA and cCyt are not active. This indicates that T cells are more sensitive to ligands presented by membranes than ligands in solution. A likely explanation for this effect is that in the confined space between the T cell membrane and that of the CHO or lipid bilayer, there will be a higher concentration

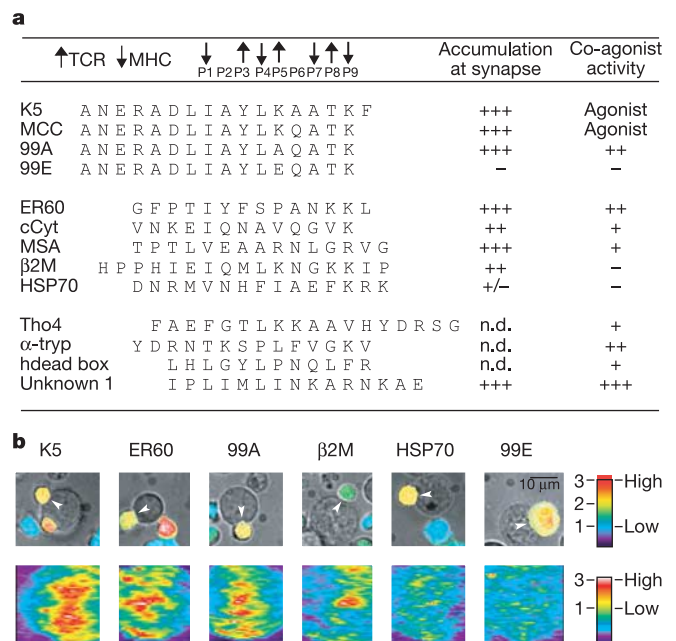


Figure 1 Self-peptide–MHCs accumulate at the T cell–APC interface. **a**, Peptide sequences are aligned according to their four anchor positions (P1, P4, P6 and P9, indicated by down arrows). Arrowheads indicate residues that affect interaction with either TCR (up arrow) or MHC (down arrow)¹⁰. Also shown is the pMHC accumulation and T cell activation score for each peptide, assayed using soluble pMHC dimers or CHO-gpi-I-E^k cells (Figs 2, 3 and Supplementary Figs S4 and S5). n.d., not determined.

b, Representative illustrations of the accumulation of pMHC complexes at the T cell–APC interface. APCs were pulsed with 100 μ M peptide before adding 5C.C7 T cells (which for visualization of endogenous peptide included 10 μ M unlabelled K5 agonist). The upper panels show differential interference contrast (DIC) images overlaid with Ca²⁺ ratio images obtained with fura-2 (340/380-nm excitations) and the lower panels show the ‘en face’ view of a 3D reconstruction of Cy3-labelled pMHC fluorescence at the T cell–APC interface. The intensity values for Ca²⁺ ratio and Cy3 accumulation ratio (synapse/membrane) are illustrated by a false colour scale, indicated by the colour bars.

We also used the CHO and lipid bilayer systems to test three additional endogenous ligands: α 1-antitrypsin (α -tryp), human dead box protein (hdead box) and Tho complex 4 (Tho4) (refs 11, 12). We found that the α -tryp ligand contributes strongly to stimulation by agonist ligand, whereas the hdead box and Tho4

ligands contribute only weakly (Fig. 1a and Supplementary Fig. S5). Taken together, these data show that the contribution of self-peptides to T cell activation observed with pMHC dimers is not an artefact of covalent linkage. We are also able to provide an explanation for the previous observation that cells bearing GPI-linked I-E^k molecules are less effective at stimulating T cells than cells expressing the same density of native molecules¹⁸. Here we

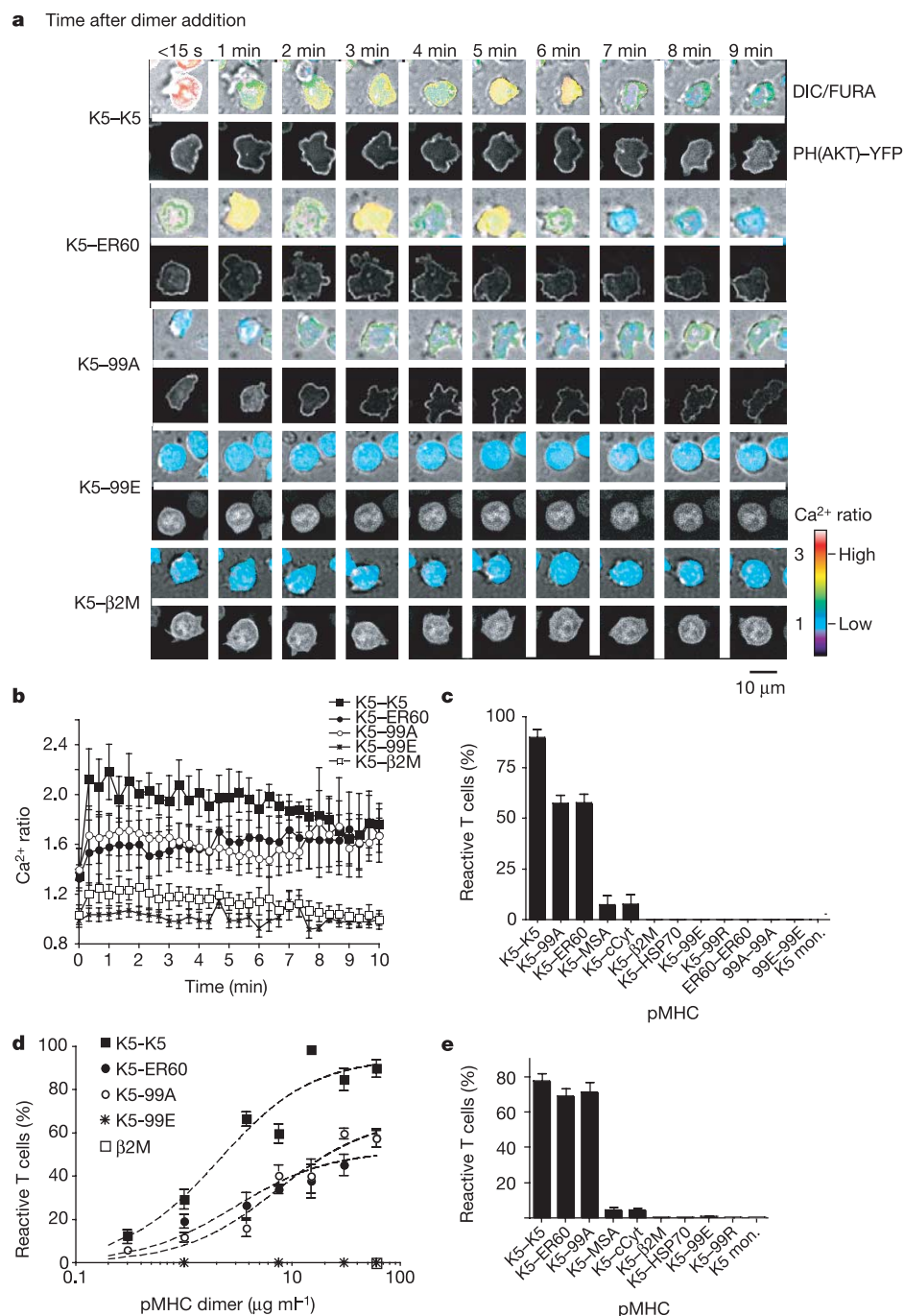


Figure 2 Soluble heterodimers of agonist and endogenous pMHCs can stimulate T cells. **a**, 5C.C7 TCR transgenic T cells expressing PH(Akt)–YFP were loaded with fura-2 and mixed with the indicated soluble pMHC dimers at 50 $\mu\text{g ml}^{-1}$. The upper panels in each horizontal pair show differential interference contrast (DIC) images overlaid with Ca^{2+} ratio images obtained with fura-2 (340/380 nm) in a false colour scale as indicated below. The lower panels show localization of PH(Akt)–YFP in T cells. **b**, Fura-2 ratios (340/380 nm) were measured every 20 s for 20 T cells in a randomly chosen microscope field (fura ratio $\pm$ s.d., $n = 20$) and the average Ca^{2+} ratio was plotted as function of time.

c. Dimers and heterodimers were added to 5C.C7 T cells at $50 \mu\text{g ml}^{-1}$ for the molecules indicated except for 99A–99A, ER60–ER60, 99E–99E and the K5 monomer, which were added at $250 \mu\text{g ml}^{-1}$. Based on the ratiometric analysis of fura-2, the percentage of cells (fura ratio $\pm$ s.d., $n = 4$) with an elevated calcium signal (>1.5) was determined.

d. The dose-response was analysed as described in **c**. **e.** The percentage of cells (>150 cells, percentage of cells $\pm$ s.d., $n = 4$) with membrane-localized PH(Akt)-YFP in response to the same concentrations of reagent as in **c**.

find that loading particular endogenous peptides onto these cells restores them to wild-type potency.

With respect to the CD4 molecule, many experiments have shown that it plays a critical role in the sensitivity of helper T cell recognition^{1,8,21}. Accordingly, we found that blocking CD4 has a deleterious effect on activation by pMHC dimers (Supplementary Fig. S6a). To discern the role of CD4 more precisely, we mutated residues Glu 137 and Thr 140 on the I-E^k β -chain to alanine^{22,23}, which abrogates CD4 activity. We made dimers using combinations of mutated (prefix m) and non-mutated (no prefix) I-E^k molecules bearing the ER60 or K5 peptides (illustrated in Supplementary Fig. S6b) and characterized their activation potential. We find that T cell activation with the K5-mER60 heterodimer is indistinguishable from that induced by the K5-ER60 heterodimer; in contrast, activation by constructs with the mutations on the agonist ligand is significantly impaired (Fig. 4a and Supplementary Fig. S6c). We then constructed dimers with the mutated CD4 binding site on both I-E^k molecules. T cell activation is almost completely abolished under these conditions (Fig. 4a and Supplementary Fig. S6), and only at high concentrations of dimer can trace Ca²⁺ fluxes be observed (data not shown). Similar effects were observed using PI(3)K activity as a marker for T cell activation (Supplementary Fig. S6d, e). We also performed these experiments using 99A (null) and K5 (agonist) ligands, and obtained very similar results (Supplementary Fig. S7).

In addition, we analysed the role of CD4 binding to agonist and endogenous peptide-MHC ligands on lipid bilayers. As in the experiments discussed above, we find no effect on T cell activation when the CD4-binding site on the endogenous ligand is mutated, compared with a major defect in T cell activation when this same site is mutated on the agonist ligand (Fig. 4b).

It is generally accepted that almost every $\alpha\beta$ TCR-bearing T cell that emerges from the thymus or is able to survive in the periphery has been selected to interact with particular endogenous peptide-MHC complexes. Many of these peptides are displayed on antigen-presenting cells in the periphery, indicating that there is a ready supply of weak self-peptides to assist in agonist stimulation. It is interesting to note that in our survey of the endogenous peptide-MHCs, accumulation at the synapse is not synonymous with co-agonist activity. Five out of the six peptides tested accumulate at the synapse, but of these, only two support robust activation, while two are weaker and two have no activity. We conclude that TCR-mediated accumulation is relatively peptide-independent (perhaps mediated just by MHC binding²⁴) and only fails when there are steric or charge constraints (for example, Glu at position 5 in MCC 99E or the bulky Phe residue at position 8 in HSP70). In contrast, the requirements for stimulation seem more demanding. However, the fact that several of the endogenous peptides analysed here have co-agonist activity suggests that there will be many more among the hundreds or thousands displayed by I-E^k molecules.

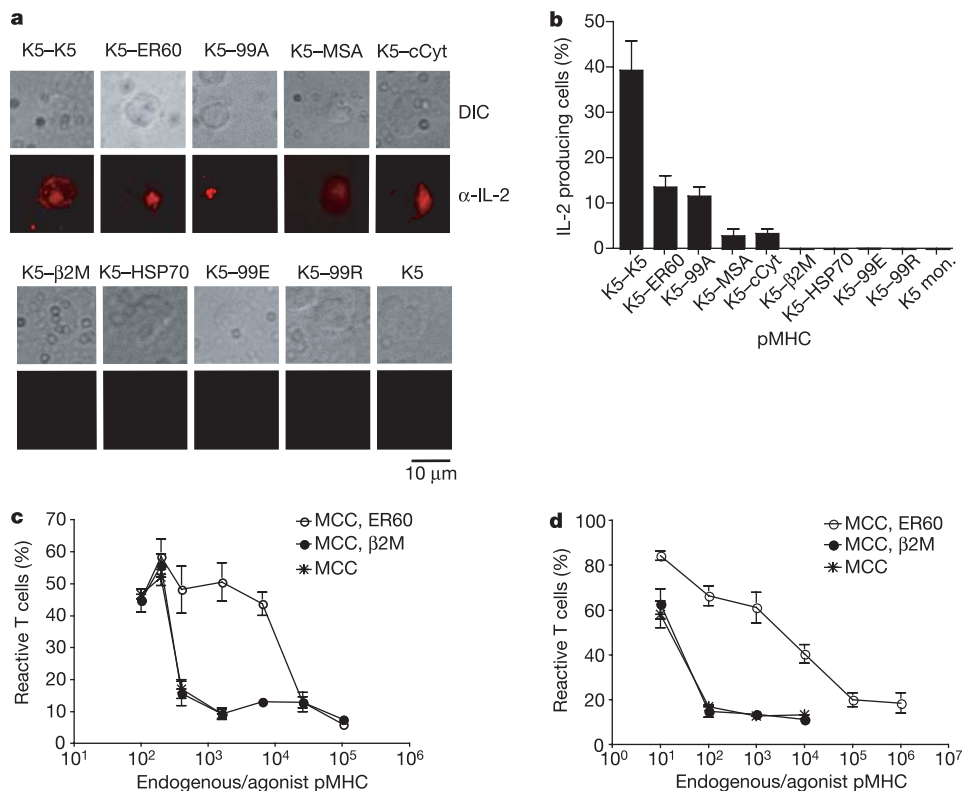


Figure 3 Soluble heterodimers can trigger IL-2 production and stimulation by endogenous and agonist pMHC on cell surfaces and lipid bilayers. **a**, Intracellular IL-2 staining. 5C.C7 TCR transgenic T cells were incubated with the indicated soluble pMHC dimers at 50 μ g ml⁻¹ at 37 °C, 5% CO₂ for 4 h. After fixation and permeabilization, IL-2 was visualized using a Cy3-labelled donkey anti-rat IgG antibody (red). The upper panels show differential interference contrast (DIC) images and lower panels show epifluorescent midplane images of Cy3. **b**, Quantification of IL-2 producing cells stained as described in **a** (>100 cells, percentage of cells $\pm$ s.d., $n = 4$). **c**, **d**, APCs or bilayers were mixed with 5C.C7 TCR transgenic T cells loaded with fura-2. The percentage of cells (>150 cells,

percentage of cells $\pm$ s.d., $n = 5$) with an elevated calcium signal (>1.5) was determined using ratiometric analysis. **c**, CHO-gpi-I-E^k APCs pulsed with a 100 μ M mixture containing the indicated ratio of endogenous peptide to agonist peptide (MCC). For the MCC control (no endogenous peptide), cells were pulsed with the same final concentration of MCC as present in the endogenous/agonist mixtures. **d**, Lipid bilayers containing B7-1 and ICAM-1 pulsed with a 0.08 μ M mixture containing the indicated ratio of endogenous-I-E^k to MCC-I-E^k or, in the case of the MCC control, the ratio of MCC-I-E^k to 'empty' MHC molecules.

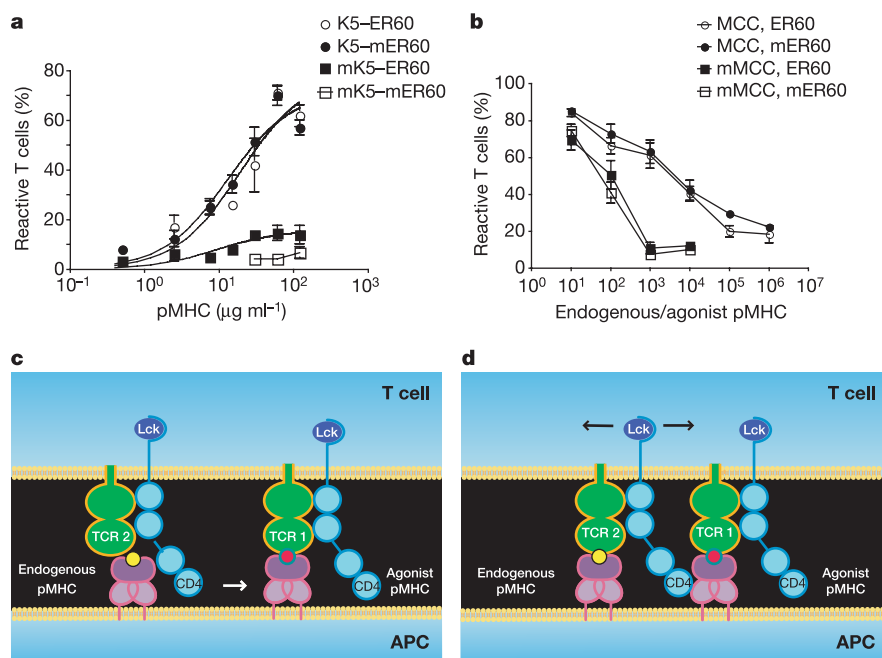


Figure 4 T cell sensitivity to mutations affecting CD4 binding and a modified pseudodimer model of activation. **a, b**, 5C.C7 T cells were loaded with fura-2 and a dose-response curve was determined for the percentage of cells with elevated Ca^{2+} signalling (340/380 nm; >150 cells; percentage of cells $\pm$ s.d., $n = 5$) when the T cells were mixed with the indicated pMHC dimers (**a**) or with lipid bilayers containing B7-1, ICAM-1 and combinations of the indicated proteins (**b**). **c, d**, A modified pseudodimer model, in

which the binding of a TCR (TCR1) to agonist pMHC (red circle) leads to the recruitment of a second TCR (TCR2), mediated by a CD4 molecule associated with it. TCR2 binds to an endogenous pMHC (yellow circle), resulting in a five-member complex stable enough to initiate activation through the tyrosine kinase Lck, phosphorylating one or both CD3 clusters (not shown).

Although the data presented here deal only with CD4^{+} T cells, we predict that CD8^{+} T cells, which recognize peptides bound to class I MHC molecules, also use self-peptide-MHC complexes in their responses. Recent peptide labelling experiments² show that two different cytotoxic T cells have exactly the same ligand sensitivity as the two helper T cell lines characterized by Irvine *et al.*¹. There may still be major differences in this respect, as some investigators find that monomeric peptide-MHC complexes can activate CD8^{+} T cells²⁵, and others find that endogenous peptide-MHC complexes play no discernable role in the response to a foreign peptide⁹. The first discrepancy is controversial²⁶, and the second may relate to different culture conditions. Given that our results apply to at least CD4^{+} T cells, how might activation be triggered? We propose a modified form of our earlier 'pseudodimer' model^{11,15} (Fig. 4c, d). In this model, the binding of a TCR (TCR1) to an agonist ligand creates a 'hotspot' for activation stable enough to recruit TCR-CD4 complexes (TCR 2). TCR2 then binds to a co-agonist endogenous pMHC and the resulting pseudodimer (TCR1/agonist peptide/TCR2/endogenous peptide/CD4) triggers an activation cascade emanating from the key protein kinase Lck carried by the recruited CD4. We have previously shown that Lck is rate-limiting with respect to early T cell activation¹⁵. We predict that the agonist ligand induces a significant conformational change to TCR1 that allows the subsequent recruitment of TCR-CD4 to occur. This may involve a twisting or piston-like motion of the TCR²⁷⁻²⁹, which may in turn facilitate CD4 binding, CD3 phosphorylation, or both.

In conclusion, we show here that T cells can be stimulated by heterodimers containing one strong and one very weak ligand. These data indicate a direct role for both endogenous peptide-MHC complexes and CD4 in triggering T cell activation. It also provides a mechanistic basis for the suggestion that the purpose of thymic selection is not only to select a TCR repertoire that can recognize foreign peptide-MHC complexes, but one that

can also use self-peptide-MHC complexes to achieve maximal sensitivity⁷. □

Methods

Constructs, peptides, mice, cells and antibodies

The genes for I-E^k α - and β -chain (including the connecting peptide region of the I-E^k β -chain (189-198) and an Ala-Cys sequence for crosslinking) were generated by polymerase chain reaction (PCR) using the plasmid pBJ1 neo/I-E^k $\alpha\beta$ (ref. 18) as source of target DNA. For constructs to be used for refolding with null and endogenous peptides (Fig. 1), the genes were fused in-frame with T7 and His-tag sequences for subsequent purification (PET28a(+), Novagen). For constructs to be used for refolding with MCC and K5 peptide (Fig. 1), the genes were cloned into a modified version of the pET28a(+) vector without the T7 sequence (a gift from E.J. Adams). For expression in insect cells, the MCC 88-103 sequence was covalently linked to the I-E^k β -chain as previously described³⁰. The E137A and T140A double mutant was introduced by site-directed mutagenesis. The 5C.C7 TCR $\alpha\beta$ transgenic mice (on a B10A background) were purchased from Jackson Laboratories. 5C.C7 blasts were generated, maintained and retrovirally infected to express PH(Akt)-YFP essentially as described¹⁷. The B cell lymphoma cell line CH27 was used as a source of APCs. For CD4-blocking studies, a monoclonal anti-CD4 antibody (GK1.5, Pharmingen) was pre-incubated with the cells for 30 min and was present during the experiment at $20 \mu\text{g ml}^{-1}$.

Crosslinking and purification of soluble pMHC complexes

Correctly folded pMHC molecules were produced essentially as described²⁹. Peptide-based crosslinking reagents (Schafer-N) were constructed and purified essentially as described⁴. Crosslinking of pMHC through the β -chain of the I-E^k molecule was done as previously described⁴. Crosslinking was verified by SDS-PAGE analysis, and for K5-K5 dimers the mixture was subjected to size-exclusion chromatography (Superdex 200, Pharmacia). For mixed dimers (K5-99A, K5-99E, K5-endogenous and K5-99A CD4 mutants) the mixture was subjected to two successive rounds of affinity chromatography (T7-tagged antibody agarose, Novagen) to remove any contaminating K5-K5 dimers before a final size-exclusion chromatography step (Superdex 200, Pharmacia).

Microscopy and image analysis

Microscopy and data analysis was carried out using a Zeiss Axiovert S100TV microscope/Metamorph (Universal Imaging Corporation) 3D time-lapse system using a $\times 40$ Fluor (NA 1.3) objective as described¹⁷. Before imaging, T cells were loaded with $10 \mu\text{M}$ fura-2(AM) (Molecular Probes) in medium for 30 min at room temperature and washed once in imaging medium (HBS plus 5% FCS, 0.5 mM CaCl_2 , 0.1 mM MgCl_2). For the accumulation data, APCs (at 2×10^6 cells ml^{-1}) were pulsed overnight at 37°C in 5% CO_2

in complete RPMI medium with a mixture of 10 μM unlabelled K5 peptide and 90 μM of an extended, N-terminal version of the biotinylated peptide synthesized as previously described¹. After three washes at 4 °C, cells were resuspended in 1% bovine serum albumin in PBS (Pierce) containing 0.5% sodium azide (to prevent internalization of pMHC complexes) and 100 $\mu\text{g ml}^{-1}$ Cy3-conjugated streptavidin (Sigma), incubated on ice for 3 h, then washed three times with imaging medium at 4 °C. Fura-loaded 5C.C7 T cells and labelled APCs were mixed together in imaging medium in Labtek eight-well chambered coverslips (Nunc) and imaged on a stage heated to 37 °C. As an indicator of intracellular Ca^{2+} concentration, fura-2 ratios were determined as previously described^{8,17}. In addition to fura-2 fluorescence imaging, the distribution of Cy3-labelled pMHCs was tracked to provide an indicator of pMHC accumulation during T cell–APC interactions. Cy3-images were acquired by collecting 21 z-stacked Cy3 planes (spaced 1 μm apart) at intervals of 1 min as previously described¹⁷. For time-lapse experiments using purified protein, T cells were mixed with pMHC dimer or MHC monomer in imaging medium in an FCS2 chamber (Biotechs, Inc.) and temperature was kept at 37 °C. Intracellular Ca^{2+} in 5C.C7 T cells was determined from fura-2 ratios as described above. The distribution of PH(Akt)–YFP in 5C.C7 T cells was measured by collecting 21 z-stacked PH(Akt)–YFP planes (spaced 1 μm apart) as previously described¹⁷. For intracellular IL-2 staining, 5C.C7 TCR transgenic T cells were mixed with the indicated soluble pMHC dimers at 60 $\mu\text{g ml}^{-1}$ and incubated at 37 °C, 5% CO_2 for 4 h. Intracellular staining was performed as described in Supplementary Methods and Cy3 images were acquired as described above.

For the CHO cell experiments, CHO cells expressing GPI-anchored I-E^k molecules¹⁸ were seeded in monolayers (2×10^5 cells ml^{-1}) in Labtek eight-well chambered coverslips (Nunc), placed on a stage heated to 37 °C and pulsed overnight at 37 °C in 5% CO_2 in complete RPMI medium with a mixture of MCC peptide and endogenous peptide (at a total peptide concentration of 100 μM). For the supported lipid bilayer experiments, mutated and non-mutated His-tagged I-E^k molecules with or without covalently linked MCC, ICAM-1 and B7-1 were expressed in the Pharmingen Baculovirus system in Hi-5 insect cells. Following Ni-NTA (nitriloacetic acid) purification, ion-exchange and size-exclusion chromatography, I-E^k molecules were labelled with endogenous peptides and mixed at a low density with I-E^k molecules that had been covalently linked with MCC, and incorporated together with ICAM-1 and B7-1 into Ni-NTA labelled, glass-supported bilayers essentially as described²⁰. Intracellular Ca^{2+} in 5C.C7 T cells was determined from fura-2 ratios as described above. Data analysis was performed in Metamorph 5.0 (Universal Imaging). To remove out-of-focus light, 3D fluorescence stacks were subjected to 50 iterations of a blind deconvolution algorithm (Deblur 9.2, Autoquant Imaging).

Received 19 December 2004; accepted 26 January 2005; doi:10.1038/nature03391.

Published online 20 February 2005.

- Irvine, D. J., Purbhoo, M. A., Krogsgaard, M. & Davis, M. M. Direct observation of ligand recognition by T cells. *Nature* **419**, 845–849 (2002).
- Purbhoo, M. A., Irvine, D. J., Huppa, J. B. & Davis, M. M. T cell killing does not require the formation of a stable mature immunological synapse. *Nature Immunol.* **5**, 524–530 (2004).
- Boniface, J. J. *et al.* Initiation of signal transduction through the T cell receptor requires the multivalent engagement of peptide/MHC ligands. *Immunity* **9**, 459–466 (1998); erratum **9**, 891.
- Cochran, J. R., Cameron, T. O. & Stern, L. J. The relationship of MHC-peptide binding and T cell activation probed using chemically defined MHC class II oligomers. *Immunity* **12**, 241–250 (2000).
- Tanchot, C., Lemonnier, F. A., Perarnau, B., Freitas, A. A. & Rocha, B. Differential requirements for survival and proliferation of CD8 naive or memory T cells. *Science* **276**, 2057–2062 (1997).
- Ernst, B., Lee, D. S., Chang, J. M., Sprent, J. & Surh, C. D. The peptide ligands mediating positive selection in the thymus control T cell survival and homeostatic proliferation in the periphery. *Immunity* **11**, 173–181 (1999).
- Stefanova, I., Dorfman, J. R. & Germain, R. N. Self-recognition promotes the foreign antigen sensitivity of naive T lymphocytes. *Nature* **420**, 429–434 (2002).
- Wulfig, C. *et al.* Costimulation and endogenous MHC ligands contribute to T cell recognition. *Nature Immunol.* **3**, 42–47 (2002).
- Sporri, R. & Reis e Sousa, C. Self peptide/MHC class I complexes have a negligible effect on the response of some CD8+ T cells to foreign antigen. *Eur. J. Immunol.* **32**, 3161–3170 (2002).
- Davis, M. M. *et al.* Ligand recognition by $\alpha\beta$ T cell receptors. *Annu. Rev. Immunol.* **16**, 523–544 (1998).
- Marrack, P., Ignatowicz, L., Kappler, J. W., Boymel, J. & Freed, J. H. Comparison of peptides bound to spleen and thymus class II. *J. Exp. Med.* **178**, 2173–2183 (1993).
- Schild, H. *et al.* Natural ligand motifs of H-2E molecules are allele specific and illustrate homology to HLA-DR molecules. *Int. Immunol.* **7**, 1957–1965 (1995).
- Lee, S. J., Hori, Y., Groves, J. T., Dustin, M. L. & Chakraborty, A. K. Correlation of a dynamic model for immunological synapse formation with effector functions: two pathways to synapse formation. *Trends Immunol.* **23**, 492–499 (2002).
- Tsien, R. Y. Fluorescent probes of cell signaling. *Annu. Rev. Neurosci.* **12**, 227–253 (1989).
- Li, Q. J. *et al.* CD4 enhances T cell sensitivity to antigen by coordinating Lck accumulation at the immunological synapse. *Nature Immunol.* **5**, 791–799 (2004).
- Deane, J. A. & Fruman, D. A. Phosphoinositide 3-kinase: diverse roles in immune cell activation. *Annu. Rev. Immunol.* **22**, 563–598 (2004).
- Huppa, J. B., Gleimer, M., Sumen, C. & Davis, M. M. Continuous T cell receptor signaling required for synapse maintenance and full effector potential. *Nature Immunol.* **4**, 749–755 (2003).
- Wettstein, D. A., Boniface, J. J., Reay, P. A., Schild, H. & Davis, M. M. Expression of a class II major histocompatibility complex (MHC) heterodimer in a lipid-linked form with enhanced peptide/soluble MHC complex formation at low pH. *J. Exp. Med.* **174**, 219–228 (1991).
- Schild, H. *et al.* The nature of major histocompatibility complex recognition by gamma delta T cells. *Cell* **76**, 29–37 (1994).
- Sumen, C., Dustin, M. L. & Davis, M. M. T cell receptor antagonism interferes with MHC clustering and integrin patterning during immunological synapse formation. *J. Cell Biol.* **166**, 579–590 (2004).
- Miceli, M. C. & Parnes, J. R. Role of CD4 and CD8 in T cell activation and differentiation. *Adv. Immunol.* **53**, 59–122 (1993).

- Konig, R., Huang, L. Y. & Germain, R. N. MHC class II interaction with CD4 mediated by a region analogous to the MHC class I binding site for CD8. *Nature* **356**, 796–798 (1992).
- Cammarota, G. *et al.* Identification of a CD4 binding site on the β_2 domain of HLA-DR molecules. *Nature* **356**, 799–801 (1992).
- Wu, L. C., Tuot, D. S., Lyons, D. S., Garcia, K. C. & Davis, M. M. Two-step binding mechanism for T-cell receptor recognition of peptide MHC. *Nature* **418**, 552–556 (2002).
- Delon, J. *et al.* CD8 expression allows T cell signaling by monomeric peptide-MHC complexes. *Immunity* **9**, 467–473 (1998).
- Ge, Q. *et al.* Soluble peptide-MHC monomers cause activation of CD8+ T cells through transfer of the peptide to T cell MHC molecules. *Proc. Natl Acad. Sci. USA* **99**, 13729–13734 (2002).
- Aivazian, D. & Stern, L. J. Phosphorylation of T cell receptor ζ is regulated by a lipid dependent folding transition. *Nature Struct. Biol.* **7**, 1023–1026 (2000).
- Sun, Z. J., Kim, K. S., Wagner, G. & Reinherz, E. L. Mechanisms contributing to T cell receptor signaling and assembly revealed by the solution structure of an ectodomain fragment of the CD3 $\epsilon\gamma$ heterodimer. *Cell* **105**, 913–923 (2001).
- Krogsgaard, M. *et al.* Evidence that structural rearrangements and/or flexibility during TCR binding can contribute to T-cell activation. *Mol. Cell* **12**, 1367–1378 (2003).
- Fremont, D. H., Hendrickson, W. A., Marrack, P. & Kappler, J. Structures of an MHC class II molecule with covalently bound single peptides. *Science* **272**, 1001–1004 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. D. Stone and L. Stern for providing the crosslinker reagent to initiate these studies as well as for helpful discussions. We thank B. Lillemeier for Baculovirus DNA encoding His-tagged ICAM-1 and B7-1. We thank C. Schaefer-Nielsen, M. Kuhns, A. Krogsgaard and members of the Davis and Chien laboratory for helpful discussions. We thank K. C. Garcia and M. Winslow for critical reading of the manuscript and helpful discussions and analysis. We thank N. Prado and B. Smith for technical assistance. M.K. was a postdoctoral fellow of the Alfred Benzon Foundation and the Danish Medical Research Council. Q.L. and M.H. are supported by Helen Hay Whitney Foundation Fellowships. This work is supported by grants from the NIH and from the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.M.D. (mdavis@cmgm.stanford.edu).

Integral role of IRF-5 in the gene induction programme activated by Toll-like receptors

Akinori Takaoka^{1*}, Hideyuki Yanai^{1*}, Seiji Kondo^{2*}, Gordon Duncan², Hideo Negishi¹, Tatsuaki Mizutani¹, Shin-ichi Kano¹, Kenya Honda¹, Yusuke Ohba^{1,3}, Tak W. Mak² & Tadatsugu Taniguchi¹

¹Department of Immunology, Graduate School of Medicine and Faculty of Medicine, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan

²Campbell Family Institute of Breast Cancer Research, University of Toronto, 610 University Avenue, Room 7-411, Toronto, Ontario M5G 2M9, Canada

³Information and Cell Function, PRESTO, JST, Kawaguchi, Saitama 332-0012, Japan

* These authors contributed equally to this work

The activation of Toll-like receptors (TLRs) is central to innate and adaptive immunity^{1–3}. All TLRs use the adaptor MyD88 for signalling⁴, but the mechanisms underlying the MyD88-mediated gene induction programme are as yet not fully understood. Here, we demonstrate that the transcription factor IRF-5 is generally involved downstream of the TLR–MyD88 signalling pathway for gene induction of proinflammatory cytokines, such as interleukin-6 (IL-6), IL-12 and tumour-necrosis factor- α . In haematopoietic cells from mice deficient in the *Irfs* gene (*Irfs*^{−/−} mice), the induction of these cytokines by various TLR ligands is severely impaired, whereas interferon- α induction is normal. We also provide evidence that IRF-5 interacts with and is activated by MyD88 and TRAF6, and that TLR activation results in the nuclear translocation of IRF-5 to activate cytokine gene transcription.

Consistently, *Irf5*^{-/-} mice show resistance to lethal shock induced by either unmethylated DNA or lipopolysaccharide, which correlates with a marked decrease in the serum levels of proinflammatory cytokines. Thus, our study identifies IRF-5 as a new, principal downstream regulator of the TLR–MyD88 signalling pathway and a potential target of therapeutic intervention to control harmful immune responses.

The IRF family of transcription factors has gained much attention for its contribution to the regulation of the immune system^{5–7}. Thus far, little is known about the function of IRF-5 of this family, although it has gained much attention in the context of type I interferon (IFN) gene induction^{8,9}. Notably, IRF-5 messenger RNA is expressed at high levels in splenic cells (Supplementary Fig. 1a and ref. 9) and it is upregulated in conventional dendritic cells and macrophages upon activation by various TLR ligands (that is, pathogen-associated molecular patterns, PAMPs) (Supplementary Fig. 1b). Although two IRF family members, IRF-3 and IRF-7, are activated by TLR stimulation for the induction of type I IFNs^{1–3,10,11}, it is unknown whether IRF-5 participates in TLR signalling.

To clarify further whether IRF-5 is involved in TLR signalling, we generated mice deficient in the *Irf5* gene (*Irf5*^{-/-} mice) by the standard homologous recombination protocol (Fig. 1a), and its nullizygosity was confirmed by DNA and RNA blot and immunoblot analyses (Fig. 1b–d). The mice developed normally and no overt difference was observed in haematopoietic cell population (Supplementary Fig. 1c). We first examined spleen-derived

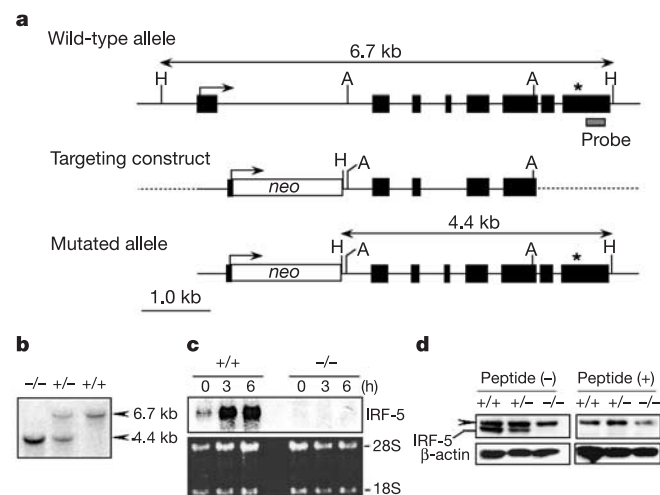


Figure 1 Generation of *Irf5*^{-/-} mice. **a**, Targeted disruption of the mouse *Irf5* gene. Organization of mouse *Irf5* gene, targeting vector and mutated allele. Filled boxes denote exons (2–9); the first (non-coding) exon is located about 4.4 kb upstream of the second exon (NCBI database Accession number NC 00072). The arrows and the asterisks indicate the translational initiator ATG and terminator TAA sequences, respectively. H, *HindIII*; A, *Apa I*. **b**, DNA blot analysis of genomic tail DNA digested with *HindIII* using the radiolabelled probe shown in **a**. The results show a single 6.7-kb band for wild type (+/+), a 4.4-kb band for homozygous (–/–) and both bands for heterozygous (+/–) mice. **c**, RNA blot analysis of RNA from poly(I:C)-stimulated MEFs. MEFs from wild-type, heterozygous and homozygous *Irf5* mutant mice were stimulated with poly(I:C) (200 µg ml⁻¹), and total RNA (10 µg) was analysed for the expression of *IRF-5* mRNA. Ethidium bromide staining after RNA transfer to the membrane is included as control (bottom panel). *IRF-5* mRNA is upregulated by poly(I:C), which activates TLR3. **d**, Immunoblot analysis of lysates from splenocytes of wild-type, heterozygous and homozygous *Irf5* mutant mice. The transferred membrane was incubated with antiserum against the *IRF-5* peptide, and were further examined for specificity by pre-incubation with excess *IRF-5* peptide. The same membrane was blotted with anti-β-actin for a loading control (bottom panel). The arrow indicates nonspecific bands. The band corresponding to the *IRF-5* protein of the right size (57 kDa) is missing in spleens from *Irf5*^{-/-} mice.

conventional dendritic cells and plasmacytoid dendritic cells for the induction of cytokines by activating TLR9 using the two TLR9 ligands CpG-B oligodeoxynucleotide (CpG-B ODN, also called 1668 or K-type ODN; refs 12, 13) and CpG-A ODN (also called D19 or D-type ODN; refs 12, 14). The induction of IL-6, IL-12 and tumour-necrosis factor-α (TNF-α) in conventional dendritic cells stimulated by CpG-B ODN or CpG-A ODN was strongly reduced in *Irf5*^{-/-} compared with wild-type cells (Fig. 2a). A similar reduction in cytokine production was observed in plasmacytoid dendritic cells from *Irf5*^{-/-} mice (Fig. 2b). In contrast, the induction of IFN-α, which is exclusive to plasmacytoid dendritic cells stimulated by CpG-A ODN^{12–15}, was normal or even enhanced in *Irf5*^{-/-} plasmacytoid dendritic cells, indicating that IRF-5 does not participate in type I IFN induction (Fig. 2b; see also Supplementary Fig. 2).

CpG-B ODN-stimulated induction of IL-6, IL-12 and TNF-α in splenic macrophages, as well as IL-6 in B cells, was also impaired in cells from *Irf5*^{-/-} mice (Fig. 2c; see also Supplementary Fig. 3a, b). Notably, impaired cytokine induction was also observed when splenic macrophages were stimulated by other TLR ligands; namely, those for TLR3 (polyinosinic-polycytidylic acid, poly(I:C))¹⁶, TLR4 (lipopolysaccharide, LPS)¹⁷, TLR5 (flagellin)¹⁸ and TLR7(8) (poly-uridylic acid, poly(U))^{19–21} (Fig. 2d). Expression of TLR mRNAs remained unaffected, and the B-cell growth and induction of immunoregulatory cell surface molecules such as CD40 and CD86 in B cells and conventional dendritic cells were also normal in the absence of IRF-5 (Supplementary Fig. 4a–c). Furthermore, loss of IRF-5 does not affect the viability of macrophages after TLR stimulation (Supplementary Fig. 4d). Therefore, IRF-5 is selectively involved in the induction of proinflammatory cytokines triggered through these and perhaps all other TLRs.

Quantitative polymerase chain reaction with reverse transcription (RT–PCR) analysis revealed that the induction was impaired at the mRNA level, suggesting the transcriptional activation of these cytokine genes by IRF-5 (Fig. 2e). Indeed, potential IRF-5-binding sites, termed the interferon-stimulated response elements (ISREs; ref. 22), were identifiable in these cytokine genes (Supplementary Fig. 5). Because all of these genes additionally contain NF-κB binding sites as well as binding sites for transcription factors activated by stress-activated protein kinase (SAPK)/JNK and p38 MAP kinases—which are also activated by TLR signalling³—we infer that IRF-5 cooperates with these transcription factors to activate these genes. We also infer that the residual mRNA induction observed in some *Irf5*^{-/-} cells is a reflection of differential contributions of these transcription factors, which would be dependent on cell types or stimuli (reviewed in ref. 23). Because we also found four potential ISREs in the *IκBζ* promoter (Supplementary Fig. 5), we examined the RNA induction in macrophages stimulated by various PAMPs and found that it was significantly reduced by the absence of IRF-5, suggesting that the *IκBζ* gene may be another target of IRF-5 (Fig. 2f). It has been shown that *IκBζ* regulates TLR signalling²⁴, but our *in vitro* (Fig. 2a–e) and *in vivo* data (see below) for *Irf5*^{-/-} mice are different, compared with those for *IκBζ*-deficient mice in terms of the induction of inflammatory cytokines. Therefore, the role of IRF-5 in *IκBζ*-dependent events will require further study.

To gain insights into the mechanism of IRF-5 activation by TLRs, we next examined whether IRF-5, which resides in the cytoplasm, forms intermolecular complexes with MyD88 and TNF receptor-associated factor 6 (TRAF6), both of which have critical roles in signalling triggered by all TLRs^{1,3}. We first expressed IRF-5 or IRF-3, each tagged with yellow fluorescent protein (YFP; hereafter referred to as YFP–IRF-5 and YFP–IRF-3), in the mouse macrophage RAW264.7 and human embryonic kidney 293T (HEK293T) cell lines, and subjected the cells to fluorescence microscopic analysis. As shown in Fig. 3a, IRF-5 was expressed in the cytoplasm, and a significant fraction of YFP–IRF-5 merged with cyan fluorescent protein (CFP)-tagged MyD88 (CFP–MyD88). Consistently, fluorescence resonance energy transfer (FRET) was selectively observed

between the co-expressed YFP-IRF-5 and CFP-MyD88, indicating that they are in direct contact with each other (Fig. 3a, b). Furthermore, IRF-5 but not IRF-3 co-immunoprecipitated with either MyD88 or TRAF6 when co-expressed in 293T cells (Fig. 3c).

We then examined MyD88- and TRAF6-dependent activation of the ISRE sequence by IRF-5 in HEK293T cells, using the p-55C1BLuc reporter gene that contains multiple ISREs. As shown in Fig. 3d, weak activation of the reporter gene by expression of IRF-5 alone was markedly enhanced either by co-expressing MyD88 and TRAF6 or by CpG-B ODN stimulation, supporting the notion that IRF-5 is fully integrated to the MyD88-TRAF6 signaling pathway. These observations are reminiscent of IRF-7—which also interacts with and is activated by these molecules to induce type I IFNs in plasmacytoid dendritic cells^{10,11}—and suggest a unique

mechanism by which the MyD88-TRAF6 pathway diverges to activate these two IRF molecules, which in turn activate distinct classes of cytokine genes.

We also examined how IRF-5 behaves upon TLR stimulation by expressing YFP-IRF-5 in RAW264.7 cells. A significant fraction of YFP-IRF-5 became detectable in the nucleus upon CpG-B ODN or LPS stimulation, supporting the view that activation of TLR9 and TLR4 invokes nuclear translocation of IRF-5 (Fig. 4a; see also Supplementary Fig. 6a, b). On the other hand, CFP-MyD88, similarly expressed in these cells, was not detectable in the nuclei even after stimulation by CpG-B ODN (Fig. 4a). Furthermore, when haemagglutinin-tagged IRF-5 (HA-IRF-5) was expressed in mouse embryonic fibroblasts (MEFs), HA-IRF-5 was detected in the nucleus of wild-type MEFs but not *MyD88*^{-/-} MEFs after

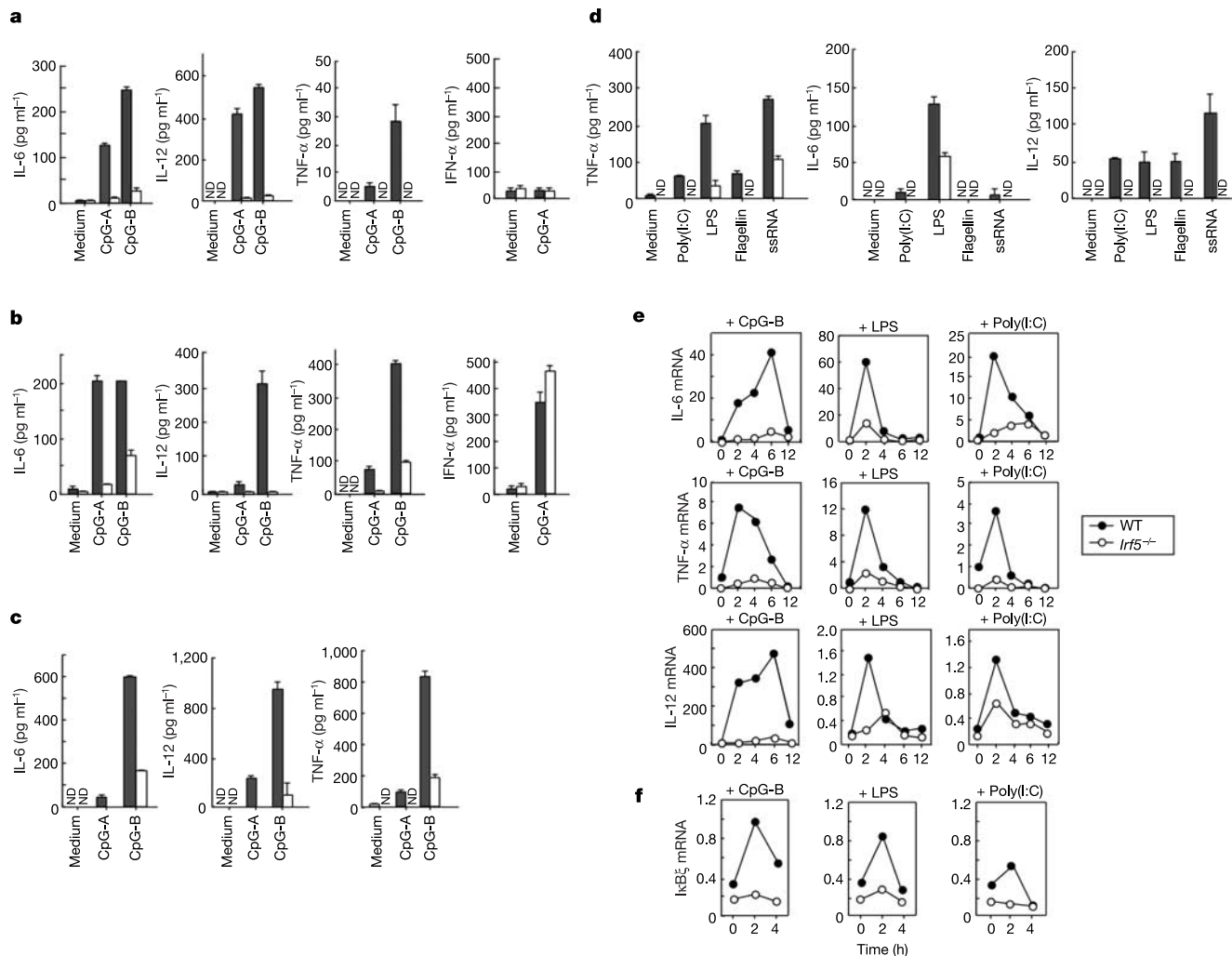


Figure 2 Impaired induction of proinflammatory cytokines and their mRNAs in haematopoietic cells from *Irif5*^{-/-} mice in response to TLR ligands. **a**, Spleen-derived conventional dendritic cells from wild-type (filled bars; **a–d**) or *Irif5*^{-/-} (open bars; **a–d**) mice were mock-stimulated (medium alone) or stimulated with 1.0 μ M CpG-A or CpG-B ODN for 24 h, and the concentrations of IL-6, IL-12p40 and TNF- α in culture supernatants were measured by ELISA. The supernatants were also measured for IFN- α induction. Similar results were obtained from three independent experiments. Data indicate mean $\pm$ standard deviation (s.d.) for **a–d**. ND, not detected. **b**, Spleen-derived plasmacytoid dendritic cells from wild-type or *Irif5*^{-/-} mice were mock-stimulated (medium alone) or stimulated with CpG-A or CpG-B ODN as described in **a**. The concentrations of the cytokines in culture supernatants were measured by ELISA. Similar results were obtained from three independent experiments. **c**, Splenic macrophages from wild-type or *Irif5*^{-/-} mice were stimulated with CpG-A (1.0 μ M) or CpG-B (1.0 μ M) ODN

for 24 h, and the concentrations of IL-6, IL-12p40 and TNF- α in culture supernatants were measured by ELISA as described in **a**. **d**, Splenic macrophages from wild-type or *Irif5*^{-/-} mice were left unstimulated or stimulated with various TLR ligands, such as those for TLR3 (poly(I:C); 100 μ g ml⁻¹), TLR4 (LPS; 10 ng ml⁻¹), TLR5 (flagellin; 10 μ g ml⁻¹) and TLR7(8) (ssRNA; 5 μ g ml⁻¹) for 24 h, and the concentration of TNF- α , IL-6 and IL-12p40 in culture supernatants was measured by ELISA. Data are representative of two independent experiments. **e**, Induction of mRNAs for proinflammatory cytokines. Splenic macrophages were prepared from wild-type and *Irif5*^{-/-} mice, and stimulated with CpG-B ODN, LPS or poly(I:C) under the same condition as **d**. Total RNAs were extracted at the indicated time, and then subjected to quantitative real-time RT-PCR analysis. **f**, Induction of *IkBε* mRNA. Splenic macrophages were stimulated by the indicated PAMPs, and *IkBε* mRNA expression was monitored as described in **e**. y-axis units for **e** and **f** are relative units.

CpG-B ODN stimulation, indicating that CpG-B-TLR9-triggered nuclear translocation of IRF-5 is indeed MyD88 dependent (Fig. 4b). In view of genetic evidence for the involvement of the IRAK kinase pathway in the TLR9-MyD88-dependent activation of IRF-7 (ref. 10), we suggest that IRF-5 is similarly activated by the IRAK kinases; this is an interesting issue to address further.

We also performed a chromatin immunoprecipitation (ChIP) assay to examine whether IRF-5 binds to one of the cytokine genes—that is, the *Il-12p40* gene, which contains a canonical ISRE element within the promoter (*Il-12p40*-ISRE; -67 to -55 from the transcription initiation site; see Supplementary Fig. 5)—by

introducing HA-IRF-5 in RAW264.7 cells. These cells also respond to CpG-B ODN, leading to the induction of these cytokines²⁵. As shown in Fig. 4c (upper panel), the PCR-amplified DNA band corresponding to *Il-12p40*-ISRE was selectively detected in the anti-HA antibody immunoprecipitate of the cells stimulated for 6 h by CpG-B ODN. The DNA band corresponding to the IFN- β promoter was not amplified in this immunoprecipitate (Fig. 4c, lower panel), an observation consistent with the fact that the IFN- β gene (*Ifnb1*) is not induced by CpG-B ODN in these cells (H.Y., unpublished data). Together, these results suggest that TLR9 signalling results in the MyD88-dependent activation of IRF-5, which in

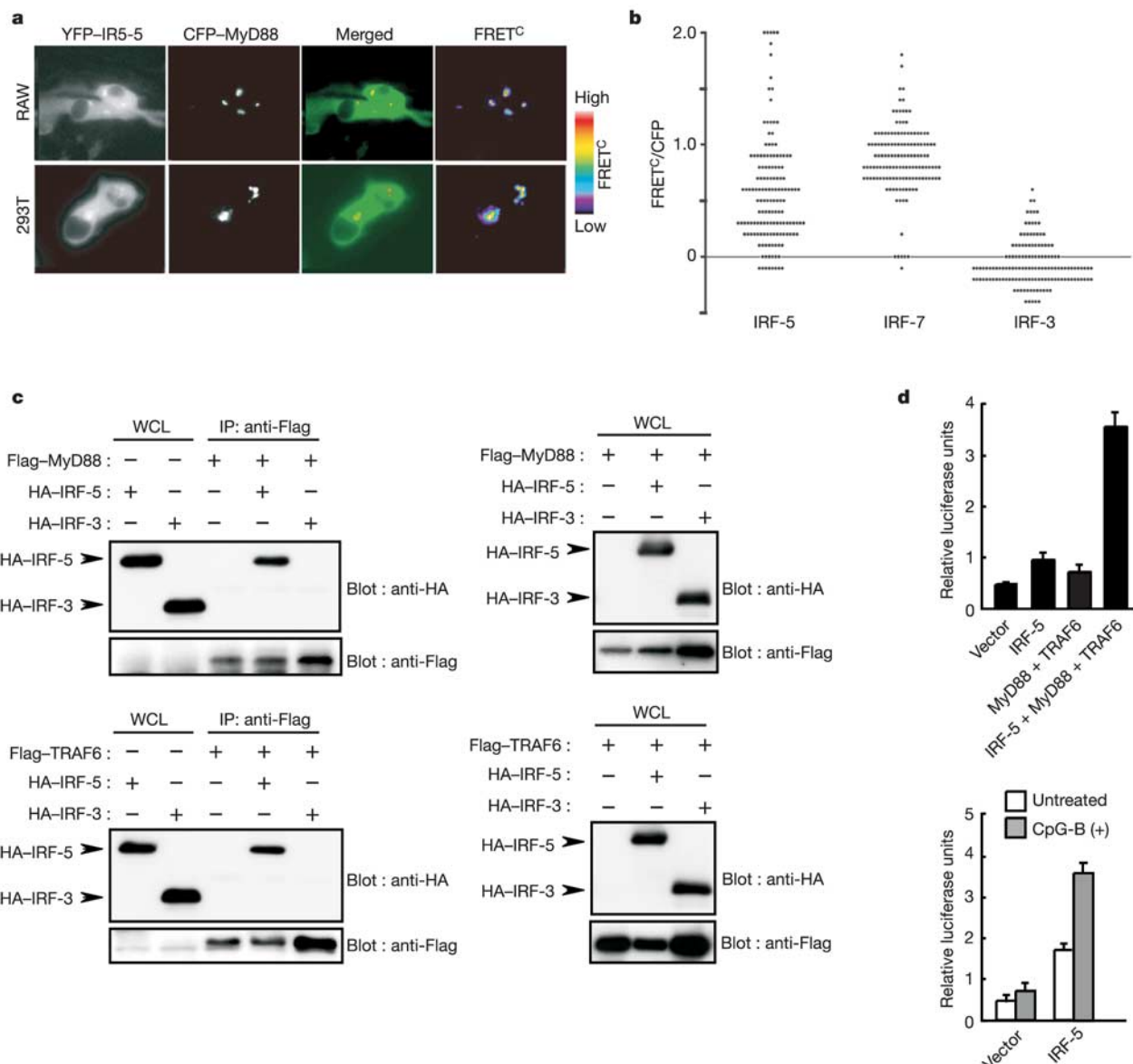


Figure 3 Interaction of IRF-5 with the MyD88 and TRAF6 adaptors for TLR signalling. **a**, Subcellular localization of IRF-5 and MyD88. YFP and CFP fluorescence, and FRET images of RAW264.7 (upper panels) and HEK293T (lower panels) cells. Cells expressing YFP-IRF-5 and CFP-MyD88 were analysed by fluorescence microscopy. Images of CFP and YFP fluorescence, merged images, and FRET^c (corrected FRET; demonstrated in pseudo-colour mode) images are shown from left to right. **b**, Intermolecular FRET analysis of binding of three IRFs to MyD88. FRET^c of the HEK293T cells co-expressing one of the indicated YFP-labelled IRFs and CFP-MyD88 was calculated, and FRET^c/CFP values were plotted as a histogram. **c**, Co-immunoprecipitation of IRF-5 or IRF-3 with MyD88 (upper panels) or TRAF6 (lower panels). HEK293T cells were transfected with the indicated

combinations of the expression vectors for HA-IRF-5, HA-IRF-3 or Flag-MyD88, and subjected to co-immunoprecipitation assay, as described in the Methods. The expression levels of these molecules were also determined by analysing the whole cell lysates (WCL) by immunoblotting. **d**, Activation of IRF-5 by MyD88 and TRAF6. A luciferase reporter construct containing multimerized ISREs (p-55C1BLuc) (50 ng) was transfected into HEK293T cells solely or together with the expression vector for the indicated combinations of IRF-5 (20 ng), MyD88 (20 ng) and TRAF6 (20 ng) (upper panel). To examine TLR9-dependent activation of ISREs by IRF-5, HEK293 cells expressing TLR9 cDNA were used (see Methods) (lower panel). After 24 h of transfection, cell lysates were extracted and subjected to luciferase assay. Data are expressed as mean $\pm$ s.d. of triplicate samples.

turn induces *IL-12p40* gene transcription. In view of the multiple ISREs within the *IL-6* and *TNF- α* genes (*Il6* and *Tnf*; Supplementary Fig. 5), we infer that these genes are similarly activated by IRF-5 through one or more of these ISREs. Identification of the specific ISREs activated by IRF-5 requires further analysis.

The TLR-dependent activation of NF- κ B, SAPK/JNK and p38 MAP kinases is dependent on the canonical MyD88-TRAF6 pathway²⁶. Because IRF-5 interacts with both MyD88 and TRAF6, we examined whether the loss of IRF-5 affects the activation of these pathways, and found that the activation of NF- κ B, p38 and JNK by CpG-B ODN was almost normal in *Irf5*^{-/-} B cells (Fig. 4d, e). We previously proposed the existence of a cytoplasmic transductional-transcriptional processor (CTTP), in which MyD88 forms a multi-molecular complex with IRF-7, TRAF6 and IRAK4 to adequately 'process' extracellular signals that in turn induce the downstream gene induction programme¹⁰. Our present study suggests that IRF-5 may be another integral member of the CTTP complex. The differential contribution of IRF-5 and IRF-7 in TLR signalling suggests the operation of a unique mechanism by which TLR-MyD88 signals are processed through the CTTP complex within a cell (for example, by spatiotemporal regulation) so as to ensure the activation of these and other transcription factors for the full-blown TLR responses. The diversification of the transcription factors,

activated downstream of MyD88 by TLR9, may occur for other TLRs.

To study the role of IRF-5 *in vivo*, we examined lethal shock induced by CpG-B ODN or LPS in D-galactosamine (D-GalN)-sensitized mice. As reported previously²⁷, wild-type mice died within 10 h after CpG-B ODN administration, with marked increases in serum IL-6, IL-12 and TNF- α concentrations; however, *Irf5*^{-/-} mice survived, and the induction of these cytokines in the serum was markedly inhibited (Fig. 5a, b). Moreover, *Irf5*^{-/-} mice also showed resistance to LPS-induced endotoxic shock (Fig. 5c, d). These results are consistent with the *in vitro* data described above, and collectively indicate the essential role of IRF-5 in the activation of TLR9 and TLR4, and possibly of all other TLRs. As expected from these data, preliminary results indicate that *Irf5*^{-/-} mice also show a deficiency in the Th1 immune response (K.H. and S. K., unpublished data).

Our study provides an insight into the mechanisms of the gene induction programme downstream of TLR signalling. Indeed, NF- κ B and MAP kinase-activated transcription factors have been extensively studied^{4,28,29}, and IRF-3 and IRF-7 are also involved in the activation of the type I interferon system^{10,11}. Our results indicate that, unlike IRF-3 and IRF-7, IRF-5 is generally involved downstream of the TLR-MyD88 signalling pathway as a master

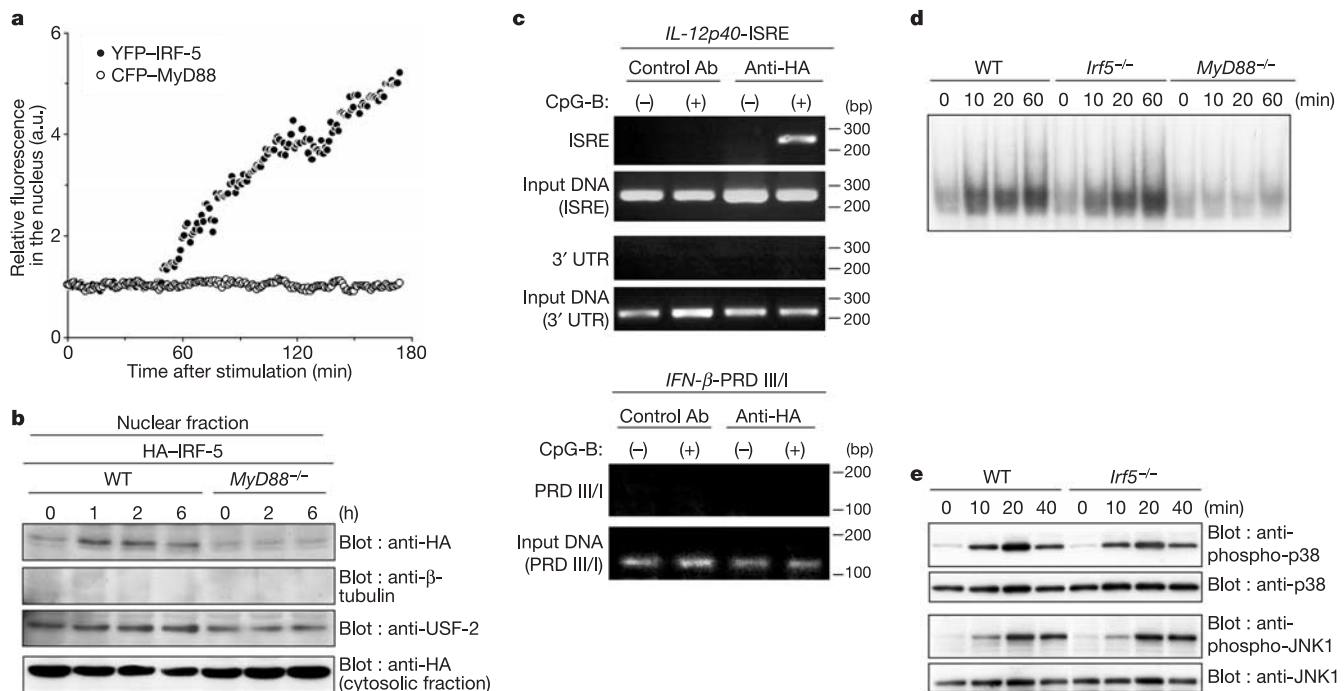


Figure 4 IRF-5 activation and binding to *IL-12p40* promoter upon TLR stimulation. **a**, Nuclear translocation of IRF-5 by CpG-B ODN stimulation. RAW264.7 cells expressing YFP-IRF-5 were imaged on a time-lapse microscope and stimulated by CpG-B ODN. Images for YFP and CFP were obtained every 1 min and CpG-B ODN was added 5 min after the start of recording. The normalized fluorescence intensities of the regions of nucleus that were identified by DIC images were plotted with time. **b**, MyD88-dependent nuclear translocation of IRF-5 in CpG-B ODN-stimulated MEFs. Wild-type or *MyD88*^{-/-} MEFs were transfected with pBabe-HA-IRF-5 via retroviral gene transfer, and treated with CpG-B ODN (1.0 μ M). Cells were fractionated and subjected to nuclear western blotting (see Methods). USF-2 is a marker for nuclear protein and β -tubulin is a marker for cytoplasmic protein. The HA-IRF-5 expression levels in the cytosolic fraction are shown at the bottom. **c**, Chromatin immunoprecipitation assay. IRF-5 binding to the endogenous ISREs in the *IL-12p40* gene was examined using RAW264.7 cells transiently expressed HA-tagged IRF-5 after stimulation with CpG-B ODN (1 μ M) for 6 h (left panel). PCR was carried out as described in Methods. Lanes 1 and 2: PCR amplification using

immunoprecipitated samples with the control antibody; lanes 3 and 4: PCR amplification of target sequences (248 bp) in immunoprecipitated chromatin fragments with the anti-HA antibody. Results of PCR amplification with primers that detect the 3' UTR of the *IL-12p40* gene are shown. The ChIP assay was similarly carried out with the primers, which allow PCR amplification of the IFN- β promoter region (right panel). *Irfb1* is not inducible by CpG-A or CpG-B ODN stimulation in these RAW264.7 cells (H.Y., unpublished data), therefore serving as a negative control. PCR amplification of the total input DNA in each sample is shown (Input DNA). The DNA size marker (in bp) is shown on the right. **d**, Activation of NF- κ B in the absence of IRF-5. Splenic B cells from wild-type, *Irf5*^{-/-} and *MyD88*^{-/-} mice were stimulated with CpG-B ODN (0.3 μ M). NF- κ B activity was assessed by EMSA. The activation of NF- κ B by poly(I:C) was also essentially normal in *Irf5*^{-/-} macrophages (A.T., unpublished data). **e**, Activation of MAP kinases in the absence of IRF-5. CpG-B ODN (0.3 μ M) was added to wild-type or *Irf5*^{-/-} splenic B cells. The activation of p38 and JNK was analysed using phospho-specific antibodies. Immunoblotting with antibodies to total p38 and JNK was carried out as a loading control.

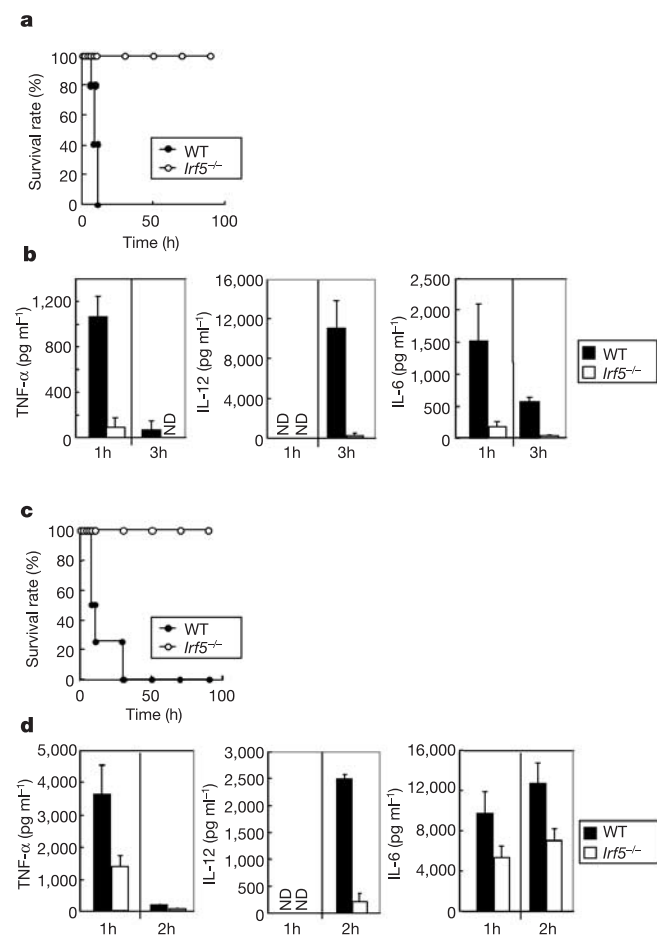


Figure 5 Resistance of *Irf5*^{-/-} mice to CpG-B ODN- or LPS-induced lethal shock. **a**, Age-matched wild-type ($n = 5$) and *Irf5*^{-/-} ($n = 4$) mice were intraperitoneally injected with CpG-B ODN (5 nmol) and D-galactosamine (D-GalN; 20 mg). Survival was monitored for 4 d. **b**, Sera of CpG-B ODN-injected mice were taken at 1 h and 3 h after injection. Serum concentrations of TNF-α, IL-12p40 and IL-6 were measured by ELISA. Results show mean $\pm$ s.d. of serum samples from three mice. **c**, Age-matched wild-type ($n = 4$) and *Irf5*^{-/-} ($n = 3$) mice were intraperitoneally injected with LPS (0.7 μg) and D-GalN (8 mg). Survival was monitored for 4 d. **d**, Sera of the LPS-injected mice were taken at 1 h and 2 h after injection as described in **b**.

transcription factor in the activation of genes for inflammatory cytokines, and possibly other genes, by various pathogen-derived molecules. Therefore, our results suggest that IRF-5 is a potential target for therapeutic interventions aimed at controlling harmful immune responses.

Methods

Generation of *Irf5*^{-/-} mice

Genomic DNA containing the *Irf5* gene was isolated from a 129J mouse genomic library. The targeting vector was constructed by replacing a 2.1-kilobase (kb) fragment encoding a part of the DNA-binding domain (corresponding to exon 2) with a neomycin-resistance gene cassette (*neo*). The *Irf5* gene-targeting vector was transfected into embryonic stem cells (E14K). Neomycin-resistant colonies were selected and screened by PCR and Southern blot analysis. Two homologous recombinants were microinjected into C57BL/6 blastocysts. Chimaeric mice were mated with C57BL/6 female mice, and heterozygous F₁ progenies were inter-crossed to obtain *Irf5*^{-/-} mice. Mice from these independent clones displayed identical phenotypes.

Other mice and reagents

MyD88^{-/-} mice were provided by S. Akira. All mice were maintained under specific pathogen-free conditions in the animal facility of the University of Tokyo and used after backcrossing with C57BL/6 mice at least seven times. Poly(U) (designated as single-stranded RNA (ssRNA) in Fig. 2d) and LPS were purchased from Sigma. The complex of poly(U) with DOTAP (Roche Diagnostics) was prepared according to the manufacturer's

instruction. Poly(I:C) was purchased from Amersham. CpG-A ODN, CpG-B ODN and all other synthetic oligodeoxynucleotides were purchased from Hokkaido System Science. Purified flagellin from *Listeria monocytogenes* was provided by A. Aderem.

Preparation of dendritic cells, macrophages and B cells

Spleens were treated with 1 mg ml⁻¹ collagenase A (Roche Biochemicals) and 20 mM EDTA, and subjected to the negative selection of T and B cells using anti-CD5 and anti-CD19 antibodies (BD Biosciences) and anti-rat IgG-coated Dynabeads (Dyna). B220⁺/CD11c⁺ conventional dendritic cells and B220⁺/CD11c⁺ plasmacytoid dendritic cells were sorted using FACS Diva (BD Bioscience). Splenic macrophages were collected using a MACS column with CD11b MicroBeads (Miltenyi Biotec) and splenic B cells were prepared by negative selection using a B-cell isolation kit according to the manufacturer's protocol (Miltenyi Biotec), and cultured in RPMI1640 (Invitrogen) supplemented with 10% fetal bovine serum.

Measurement of cytokine production

Conventional and plasmacytoid dendritic cells (2×10^5 cells ml⁻¹), B cells (5×10^5 cells ml⁻¹) or macrophages (7×10^5 cells ml⁻¹) were seeded into 96-well plates and cultured with the indicated stimuli for 24 h. The concentrations of IL-12p40, IL-6, TNF-α and IFN-α in the cultured supernatants were measured by enzyme-linked immunosorbent assay (ELISA). ELISA kits for mouse IL-12p40, IL-6 and TNF-α were obtained from R&D Systems. An ELISA kit for mouse IFN-α was purchased from PBL Biomedical Laboratories. In the case of macrophages, the supernatants cultured in the presence of 30 Units ml⁻¹ IFN-γ were used for the measurement of the concentrations of IL-12p40.

RNA analysis

RNA extraction and RT-PCR and RNA blot analysis were performed as described previously³⁰. The complementary DNA probe for IRF-5 corresponds to the 278–739 fragment. Quantitative real-time RT-PCR analysis was performed using a LightCycler and SYBRGreen system (Roche). Data were normalized by the level of β-actin expression in each sample. The primers for β-actin, IL-6, TNF-α, IL-12p40, IRF-5 and IκBα are described in Supplementary Information.

Plasmids and gene transfer

Flag-MyD88, CFP-MyD88, YFP-IRF-3, YFP-IRF-7 and HA-IRF-3 expression vectors have been described previously (see Supplementary Information). Mouse IRF-5 cDNA was obtained by RT-PCR of total RNA from MEFs, and cloned into the pCIR2 vector (Stratagene). To generate YFP-IRF-5 or HA-tagged IRF-5 expression vectors, the cDNA was cloned into the *Xho*I and *Not*I sites of the pCAGGS-YFP or pCAGGS-HA vector. pCAGGS vector or Venus, which was referred to as YFP, were provided by J. Miyazaki or A. Miyawaki, respectively. The HA-IRF-5 DNA fragment was excised from pCAGGS-HA-IRF-5 and cloned into the *Eco*RI and *Sal*I sites of the pBabe-puro retroviral expression vector. Retroviral gene transfer into MEFs was carried out as described previously³⁰. RAW264.7 cells were transfected with pCAGGS-HA-IRF-5 using the Superfect reagent (Qiagen) for ChIP assay. The expression vector for Flag-tagged TRAF6 was provided by J.-I. Inoue.

Fluorescence microscopy

Cells were cultured on glass-bottomed 35-mm tissue culture dishes (Matsunami Glass) and imaged on an inverted fluorescence microscope equipped with a cooled CCD camera, as described previously¹⁰. FRET analysis was carried out as described previously¹⁰.

Immunoprecipitation and immunoblotting

Cell lysis, co-immunoprecipitation and immunoblotting were carried out as described previously^{10,30}. Rabbit antisera against IRF-5 were raised by immunization of the peptide PPSDIPSDKQ, which corresponds to a region in exon 6 of the *Irf5* gene. This peptide was also used in the competitive assay shown in Fig. 1d. Antibodies against the following proteins were purchased as indicated: HA (Roche), Flag (Sigma-Aldrich), phospho-SAPK/JNK, SAPK/JNK, phospho-p38 MAPK and p38 MAPK (Cell Signalling Technology).

Chromatin immunoprecipitation assay

The ChIP assay was performed as described previously³⁰. The specific antibodies used for immunoprecipitations were the anti-HA antibody (Roche) and the anti-Lck antibody as an isotype control (Upstate biotechnology). The purified DNA was analysed by PCR amplification using primers that detect sequences containing IL-12p40-ISRE (nucleotide number -67 to -55): 5'-ACCCCGAAGTCATTTCCTCT-3' (sense) and 5'-ACCCACTG TTCCTCTGCT-3' (antisense); and with a pair of negative control primers that recognize the DNA sequence (+12863 to +13906) of the 3' untranslated region (UTR) of the IL-12p40 gene: 5'-GATGCAACGTTGGAAGGA-3' (sense) and 5'-TTCAACAGCATAA GGCCAAG-3' (antisense). The region of the IFN-β promoter (nucleotide number -74 to +48) was amplified by using the following primers: 5'-GGGAGAACTGAAAGTGGG AAA-3' (sense) and 5'-ACCTGCAAGATGAGGCAAG-3' (antisense). The PCR products were visualized on an ethidium bromide gel.

Reporter assay

HEK293T cells seeded on 12-well plates were transiently co-transfected using FuGene 6 reagent (Roche) with 50 ng of the reporter plasmid p-55C1BLuc (provided by T. Fujita; see Supplementary Information) and the indicated amounts of expression plasmids for IRF-5, MyD88 and TRAF6. For the CpG-B ODN stimulation experiment, HEK293 cells expressing TLR9 (InvivoGen) were transfected with 50 ng of the reporter plasmid and

100 ng of pCAGGS-HA-IRF-5, followed by CpG-B ODN stimulation for 12 h. Cells were lysed 24 h after transfection and luciferase activity was measured using the Dual-luciferase reporter assay system (Promega) as described previously¹⁰.

Additional assays

Nuclear western blotting and electrophoretic mobility shift assays (EMSA) were performed as previously described (see Supplementary Information).

Received 26 November 2004; accepted 4 January 2005; doi:10.1038/nature03308.

Published online 23 January 2005.

1. Medzhitov, R. Toll-like receptors and innate immunity. *Nature Rev. Immunol.* **1**, 135–145 (2001).
2. O'Neill, L. A., Fitzgerald, K. A. & Bowie, A. G. The Toll-IL-1 receptor adaptor family grows to five members. *Trends Immunol.* **24**, 286–290 (2003).
3. Akira, S. & Takeda, K. Toll-like receptor signalling. *Nature Rev. Immunol.* **4**, 499–511 (2004).
4. Medzhitov, R. *et al.* MyD88 is an adaptor protein in the hToll/IL-1 receptor family signaling pathways. *Mol. Cell* **2**, 253–258 (1998).
5. Taniguchi, T., Ogasawara, K., Takaoka, A. & Tanaka, N. IRF family of transcription factors as regulators of host defense. *Annu. Rev. Immunol.* **19**, 623–655 (2001).
6. Levy, D. E., Marie, I. & Prakash, A. Ringing the interferon alarm: differential regulation of gene expression at the interface between innate and adaptive immunity. *Curr. Opin. Immunol.* **15**, 52–58 (2003).
7. Barnes, B., Lubyova, B. & Pitha, P. M. On the role of IRF in host defense. *J. Interferon Cytokine Res.* **22**, 59–71 (2002).
8. Lin, R., Yang, L., Arguello, M., Penafuerte, C. & Hiscott, J. A CRM1-dependent nuclear export pathway is involved in the regulation of IRF-5 subcellular localization. *J. Biol. Chem.* (in the press).
9. Barnes, B. J., Moore, P. A. & Pitha, P. M. Virus-specific activation of a novel interferon regulatory factor, IRF-5, results in the induction of distinct *interferon α* genes. *J. Biol. Chem.* **276**, 23382–23390 (2001).
10. Honda, K. *et al.* Role of a transductional-transcriptional processor complex involving MyD88 and IRF-7 in Toll-like receptor signaling. *Proc. Natl Acad. Sci. USA* **101**, 15416–15421 (2004).
11. Kawai, T. *et al.* Interferon- α induction through Toll-like receptors involves a direct interaction of IRF7 with MyD88 and TRAF6. *Nature Immunol.* **5**, 1061–1068 (2004).
12. Verthelyi, D., Ishii, K. J., Gursel, M., Takeshita, F. & Klinman, D. M. Human peripheral blood cells differentially recognize and respond to two distinct CPG motifs. *J. Immunol.* **166**, 2372–2377 (2001).
13. Krieg, A. M. CpG motifs in bacterial DNA and their immune effects. *Annu. Rev. Immunol.* **20**, 709–760 (2002).
14. Krug, A. *et al.* Identification of CpG oligonucleotide sequences with high induction of IFN- α/β in plasmacytoid dendritic cells. *Eur. J. Immunol.* **31**, 2154–2163 (2001).
15. Colonna, M., Trinchieri, G. & Liu, Y. J. Plasmacytoid dendritic cells in immunity. *Nature Immunol.* **5**, 1219–1226 (2004).
16. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF- κ B by Toll-like receptor 3. *Nature* **413**, 732–738 (2001).
17. Poltorak, A. *et al.* Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in *Tlr4* gene. *Science* **282**, 2085–2088 (1998).

18. Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
19. Diebold, S. S., Kaisho, T., Hemmi, H., Akira, S. & Reis e Sousa, C. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Science* **303**, 1529–1531 (2004).
20. Heil, F. *et al.* Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. *Science* **303**, 1526–1529 (2004).
21. Lund, J. M. *et al.* Recognition of single-stranded RNA viruses by Toll-like receptor 7. *Proc. Natl Acad. Sci. USA* **101**, 5598–5603 (2004).
22. Darnell, J. E. Jr, Kerr, I. M. & Stark, G. R. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* **264**, 1415–1421 (1994).
23. Alvarez, M., Rhodes, S. J. & Bidwell, J. P. Context-dependent transcription: all politics is local. *Gene* **313**, 43–57 (2003).
24. Yamamoto, M. *et al.* Regulation of Toll/IL-1-receptor-mediated gene expression by the inducible nuclear protein I κ B ζ . *Nature* **430**, 218–222 (2004).
25. Cowdery, J. S., Boerth, N. J., Norian, L. A., Myung, P. S. & Koretzky, G. A. Differential regulation of the *IL-12 p40* promoter and of *p40* secretion by CpG DNA and lipopolysaccharide. *J. Immunol.* **162**, 6770–6775 (1999).
26. Hacker, H. *et al.* Immune cell activation by bacterial CpG-DNA through myeloid differentiation marker 88 and tumor necrosis factor receptor-associated factor (TRAF)6. *J. Exp. Med.* **192**, 595–600 (2000).
27. Sparwasser, T. *et al.* Macrophages sense pathogens via DNA motifs: induction of tumor necrosis factor- α -mediated shock. *Eur. J. Immunol.* **27**, 1671–1679 (1997).
28. Cao, Z., Xiong, J., Takeuchi, M., Kurama, T. & Goeddel, D. V. TRAF6 is a signal transducer for interleukin-1. *Nature* **383**, 443–446 (1996).
29. Wesche, H., Henzel, W. J., Shillinglaw, W., Li, S. & Cao, Z. MyD88: an adapter that recruits IRAK to the IL-1 receptor complex. *Immunity* **7**, 837–847 (1997).
30. Takaoka, A. *et al.* Integration of interferon- α/β signalling to p53 responses in tumour suppression and antiviral defence. *Nature* **424**, 516–523 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Vilcek, R. Flavell, H. Rosen and N. Fujiwara for advice, and M. Shishido for technical assistance. This work was supported in part by a grant for Advanced Research on Cancer and a Grant-In-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan, Uehara Memorial Foundation, Mitsukoshi Foundation, the Princess Takamatsu Cancer Research Fund and Canadian Institute of Health Research. H.Y. is a research fellow of the Japan Society for the Promotion of Science. H.N. and S.K. were supported by an Ishidu Shun Memorial Scholarship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.W.M. (tmak@uhnres.utoronto.ca) or T.T. (tada@m.u-tokyo.ac.jp).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

European Sales Manager:

Andy Douglas (4975)

UK/ RoW/ Ireland:

Nils Moeller (4953)
Irene Vigila-Alton (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Advertising Production

Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Niamh Shields

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan

Tel + 49 89 54 90 57 11

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakatsunomachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

In the picture

A few weeks ago, face-down on a treatment table, I experienced a biomedical-imaging epiphany. While anticipating a needle's entry into a gap between two cervical vertebrae — necessary to shoot a numbing agent, then an imaging agent, then a steroid into a strategic spot in my spinal column — a profound thought flashed into my mind: this is really going to hurt.

But after the treatment began, I realized how profoundly biomedical imaging can affect scientific careers. It could, in fact, lead to a demand for scientists even greater than the bioinformatics boom in the run-up to the completion of the human genome.

In both bioinformatics and bioimaging, there are two broad classes of scientific professionals, both very much in demand — people who can make stuff and people who can use it well. But bioimaging has the potential to provide more career opportunities, because in this field those two categories of people are both broad and complex (see page 252). In my situation, chemists created the bioimaging agent, physicists and engineers designed the imaging machine, and my physician combined imaging and treatment.

Plenty of opportunities exist, spanning many disciplines: medical physicists who can make imaging isotopes; molecular biologists who can use knowledge of receptors and ligands to target imaging agents, then treatment, to specific cell types; mathematicians who can write algorithms to combine imaging data from different sources ... the list of convergences goes on.

As a patient, I was very grateful for the convergence that made my treatment possible, as well as my physician's finesse in providing it. It didn't hurt. And neither would considering how biomedical imaging can affect your career — whether in terms of creating tools or using them for research.

Paul Smaglik

Naturejobs editor



Contents

CAREERS AND RECRUITMENT

Biomedical imaging comes into focus p252

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Come together

The addition of molecular biology to the existing range of imaging technologies is creating opportunities for scientists of many disciplines. Paul Smaglik lines up the pieces.

Convergence is the watchword for a multitude of trends in biomedical imaging: imaging joined with therapy, for example, or positron emission tomography (PET) used with magnetic resonance imaging (MRI). It's also about applying methods traditionally used in hospitals to basic research, and tweaking them to improve drug development.

This scientific and technological confluence is yielding an abundance of opportunities. Engineers are needed to make different kinds of machines talk to each other. Mathematicians are needed to work out the complex algorithms for combining images of different dimensions from different machines, and to make each reagent or molecular probe work optimally with each machine. Information-technology (IT) specialists are needed to integrate floods of disparate data, then find statistically sound ways to evaluate and model them.

Another factor is that the various sectors — big and small business, government and foundation, academic and private — seem to be working symbiotically rather than cannibalizing each other, perhaps because there's not enough money and talent to go around. Training can take 20 years before people feel they've really grasped each facet (see 'Twenty-year hitch', below). And because there is not a lot of suitable training available, alliances between universities and companies, or large and small companies, may be the only way to build up the skilled workforce that's needed (see 'A squeezed pipeline', opposite).

Convergence played a role in creating the biggest player in the biomedical-imaging arena. When Wisconsin-based GE Medical Systems, which was primarily known for imaging hardware, acquired Amersham Biosciences, based in Uppsala, Sweden, GE was already combining approaches. It was using



computed tomography (CT) imaging — which emphasizes structural imaging — in conjunction with, for example, PET, which GE senior engineer Bob Armstrong describes as “all function and no structure”. Adding molecular imaging, a growing discipline in and of itself (see *Nature Med.* 11, 111–112; 2005), means that “the next thing is looking at how things metabolize” within the body, Armstrong says.

It also means that, unlike most post-merger companies, the new GE Healthcare is adding jobs. Combining GE's physicists and engineers with Amersham's biologists has created what Armstrong calls “a big gap”. Chemists — especially biochemists — will help to make up the numbers.

But that will require some serious information integration, says Dorin Comaniciu, department head for integrated data systems at Siemens Corporate Research in Princeton, New Jersey. And integrating data from different kinds of imaging techniques requires more than just computer savvy, he says.

“You cannot come only from the IT point of view,” Comaniciu says. “You have to understand the complexity of the data. You have to understand the complexity of the human body.”

INFORMATION OVERLOAD

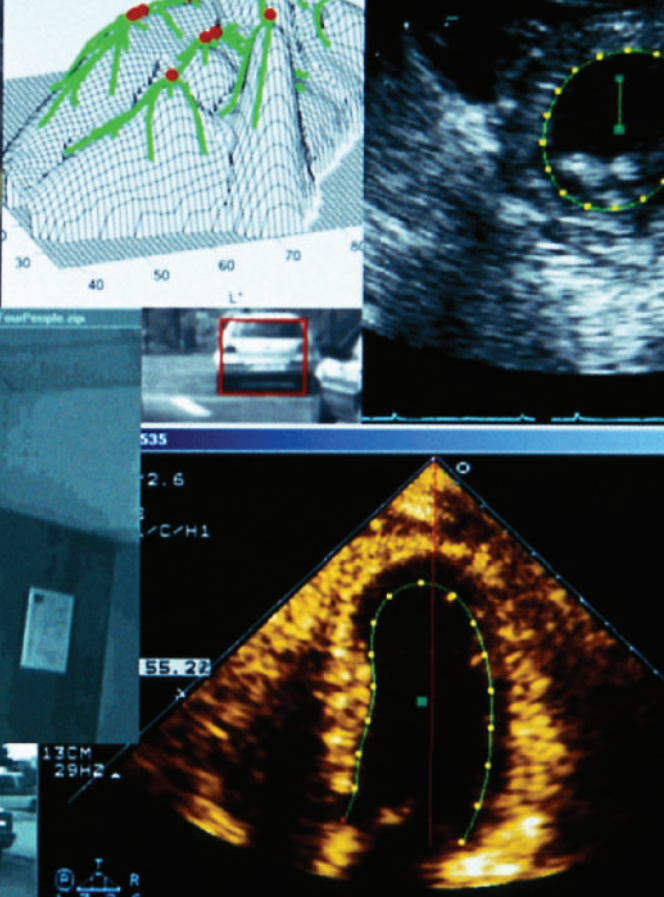
When recruiting, Comaniciu seeks computer scientists who can make that complexity easier to handle, by seeing the data from the perspectives of different users: the engineers and physicists designing the hardware, the physicians and researchers using it, and the biologists and chemists who add molecular components.

In fact, convergence has created a need for two kinds of scientists, says David Rollo, chief medical officer at Philips Medical Systems in Milpitas, California. To optimize imaging of specific diseases, Philips aims to find a specific molecular or radiological probe that best targets different disease types —

Twenty-year hitch

The good news for young scientists seeking a career in biomedical imaging is that jobs are widely available and salaries are high. The bad news is it may take 20 years of training and experience to get to that level.

Perhaps the most extreme example is one of the world's most visible bioimaging experts, Roderic Pettigrew, who heads the US National Institute of Biomedical Imaging and Bioengineering in Bethesda, Maryland. While completing a PhD in nuclear physics, Pettigrew became interested in the infant field of magnetic resonance imaging. He became curious about what problems the nascent technology could peer into, so he pursued an accelerated course of medical training available only to people who have a PhD. After two years of study with no vacation, he earned an MD, then did a medical residency. He then joined industry to understand how companies build machines, and went back to academia as a professor. Now, after working in seemingly every possible sector, Pettigrew jokes that his whole career was headed towards his current position. **P.S.**



SIEMENS

a molecular probe that binds only to a cell receptor found in just one kind of cancer, for instance. So the company needs biologists and physicists to find and develop these probes. But mathematicians are even more in demand, he says, if they can write algorithms specific to such a molecule that will produce an accurate, high-resolution picture when working with all the other imaging data.

Because of this, convergence is creating markets for newly launched companies that take existing imaging technology and add a therapeutic twist. These firms often work in partnership with the three big imaging companies: GE Healthcare, Siemens and Philips.

For example, Philips works with Kereos of St Louis,

Missouri, a biotech firm that emerged from Washington University. The university has some 35 researchers developing novel chemotherapeutics and contrast agents. These scientists act as a basic research group for Kereos, which picks and chooses the agents it wants and acts as the development arm for Philips.

IN DEMAND

An interaction between the University of Wisconsin and imaging company TomoTherapy illustrates one of the bioimaging industry's main needs — medical physicists. Thomas Mackie, professor of medical physics at the University of Wisconsin and director of research at TomoTherapy, anticipates that staff numbers at his company alone will almost double to about 475 this year, with that subspecialty in demand. Salaries and help-wanted ads reflect this demand, Mackie says, citing the American Association of Physicists in Medicine, which says the mean salary is \$125,000 and has ample job ads on its website.

Mackie says there are two ways to get into this field: attend a medical physics course endorsed by the Commission on Accreditation of Medical Physics Educational Programs; or, if you already have a PhD in a physics subspecialty such as plasma, nuclear or high-energy physics, aim for a specialized medical residency.

Doug Bakan, vice-president of business development for San Diego-based Alereon, offers a more direct route into the industry. Consider starting in chemistry, then branching out. Chemists are much in demand at the company, which makes contrast agents for imaging. People who can design and synthesize molecules are especially needed, as are analytical biochemists who look at the purity, functionality and stability of molecules.

But to excel in biomedical imaging, it's not enough just to know one's specialism. "Be good at something, then learn something about every component over time," Bakan says. Scientists heeding this advice are likely to find their career coming together. ■

Paul Smaglik is editor of *Naturejobs*.

A squeezed pipeline

Training in biomedical imaging has a pipeline problem, say many professionals in the field. The need for scientists with multidisciplinary skills has been growing, while training centres that offer all of the components are rare. And budgetary constraints in the United States — the leader in the field — may hamper opportunities at a time when demand is growing.

Thomas Mackie, professor of medical physics at the University of Wisconsin in Madison and director of research at imaging company TomoTherapy, says that the "feedstock" in the three major programmes that siphon into the field are in jeopardy. "Physics is shrinking, biomedical imaging is growing, but under a lot of pressure, and nuclear physics has been shrinking for years," Mackie says.

Michael Welch, a professor of radiology, chemistry, molecular biology and pharmacology at Washington University in St Louis, agrees that the issue is especially acute in nuclear imaging. He is concerned that the US Department of Energy, which has played a big role in funding training in nuclear medicine, had its

education budget gutted in President George W. Bush's proposed budget for 2006. And the proposed budget for the US National Institutes of Health, which offers a mentoring training programme for physicists wanting to apply their skills to the life sciences, is a flat.

There are glimmers of hope that other players will step up to the bioimaging training plate. The Howard Hughes Medical Institute in Chevy Chase, Maryland, is calling for proposals to fund interdisciplinary training, including biomedical imaging.

Industry is also showing signs of helping meet its own training needs. Siemens Corporate Research, for instance, has 100 interns alone in its facility in Princeton, New Jersey, one of five research and development spots globally.

But the US training shortage is likely to have two positive knock-on effects. First, imaging centres in other parts of the world (see Web links) will play a bigger role in training. And students who manage to get into and complete these programmes are likely to find themselves in high demand.

P.S.

Web links

HHMI-NIBIB Interfaces Program

♦ www.hhmi.org/grants/institutions/nibib.html

NIH Mentored Research Scientific Development Award

♦ grants2.nih.gov/grants/guide/pa-files/PA-00-019.html

American Association of Physicists in Medicine

♦ www.aapm.org
Commission on Accreditation of Medical Physics Educational Programs

♦ www.campep.org

A man of the theatre

All the world's a stage.

Norman Spinrad

I've been a man of the theatre for over a century. Old enough to have played Hamlet, and Richard, and any number of Henries, and Lear before my first rejuvenation — and all in my own flesh and on the stage, not as an avatar, licensed or otherwise.

You remember the theatre, don't you?

Of course you don't. There hasn't been a play produced in what, a quarter of a century. In what I call a play, the characters must be played by live actors, not software emulations of the dead greats of yesteryear, who never played the parts themselves — and certainly not by one of myself.

And not because I could not compete with these pathetic avatars. Most of the roles in realies are played by the greats of the past, but because they all died before fleshware downloading technology was available, they've all been synthesized from old footage by the entertainment conglomerates that owned the rights, and they walk mechanically through the roles like the virtual robots they are, not even the virtual ghosts of true thespians.

What is more, I was offered an emulation contract many long years ago, which I scornfully turned down. Yes, I was that good. Good enough that they were willing to pay me royalties for the use of my avatar, even though they had a free casting call from over a century of film and television. But I would not betray the theatre for any amount of money.

The theatre, they declare, is obsolete. Why would anyone pay money to watch third-rate human actors staggering around on a platform in front of a flat painted set when they can tap into full virtual realities telling the same story, if it's any good, with smell, and taste, and touch, and pleasure-centre stimulation, within a fully realized world that can be synthesized as cheaply as theatrical sets? And to reach what demographics? A few thousand people a day when the same budget creates a hit machine tapped by scores of millions!

The theatre is dead.

But the theatre must not die. Those who do not understand why have never set foot on a stage before a live audience. I may be the last man alive who has. You have not seen those shining eyes beyond the footlights riveted on you; you have not smelled the heady aroma of an expectant live audience. Yes, I know, if there was any interest in such ancient history, you could experience it in a realie. Or so you believe.

But the magic of the theatre cannot be emulated: the intimate connection between the human actor and a live audience. For



when the play succeeds, there is a collaboration between the actors and the audience — the actors and the audience live and breathe together, a community of the spirit in which the reaction of the audience influences the actors and shapes the live audience's experience itself. A positive feedback loop, as you moderns would so unromantically have it.

If the popular 'entertainments' that fill so many broadcast hours are all soulless exercises, with none of the drama acted out by fellow humans in the intimacy of a living community, life itself will be entirely reduced to virtuality. Has it not already happened in the retirement heavens, where zombies are tapped into a thousand available channels of realies, 24 hours a day? The life-support technology already exists and it only awaits a profitable business model for the entire population of the planet to dream their lives away as the solipsistic gods of their perpetual virtual heavens.

This is neither drama nor life. This is tyranny with an entertaining face. If the theatre dies, so dies the human spirit. For only a great act of theatre can reawaken it.

Antonin Artaud wrote of the Theatre of Cruelty: do not amuse your audience. Be cruel to your audience in order to seize and hold it. I shall go him one further.

The King will be parading from Buckingham Palace to the Houses of Parliament in his magnificently baroque horse-drawn carriage and in full costume, like the penultimate actor that he is, his simple performance witnessed live by throngs gathered along the way to experience the Royal Presence at first hand.

I shall wear the costume I wore when I played Othello the Moor, which no doubt will be taken for that of a Caliphate terrorist, and the sword I shall use will be a scimitar, which I shall plunge into the Royal breast before I detonate the explosives under my robes. It should bring on the long-awaited war. Thus will I remind the world of the sovereign power of an act of live theatre.

And my last line on the stage will be that declaimed in like manner by a scion of a noble theatrical family whose name yet lives for the ages. Not that of the great Edwin but that of his otherwise mediocre brother. John Wilkes Booth.

"Sic semper tyrannus!" ■

Norman Spinrad's latest novel, *Mexica*, will be published by Little Brown in November 2005. He is a novelist, screenwriter, critic and essayist. He writes the occasional song, paints the occasional painting, and does the very occasional turn as rock star. He is currently bouncing back and forth between Paris and New York.